

Efficient Training for Positive Unlabeled Learning

Emanuele Sansone, Francesco G. B. De Natale, *Senior Member, IEEE* and Zhi-Hua Zhou, *Fellow, IEEE*

Abstract—Positive unlabeled (PU) learning is useful in various practical situations, where there is a need to learn a classifier for a class of interest from an unlabeled data set, which may contain anomalies as well as samples from unknown classes. The learning task can be formulated as an optimization problem under the framework of statistical learning theory. Recent studies have theoretically analyzed its properties and generalization performance, nevertheless, little effort has been made to consider the problem of scalability, especially when large sets of unlabeled data are available. In this work we propose a novel scalable PU learning algorithm that is theoretically proven to provide the optimal solution, while showing superior computational and memory performance. Experimental evaluation confirms the theoretical evidence and shows that the proposed method can be successfully applied to a large variety of real-world problems involving PU learning.

Index Terms—Machine learning, one-class classification, positive unlabeled learning, open set recognition, kernel methods

1 INTRODUCTION

POSITIVE unlabeled learning (PU learning) refers to the task of learning a binary classifier from only positive and unlabeled data [1]. This classification problem arises in various practical situations. For example:

- Retrieval [2], where the goal is to find samples in an unlabeled data set similar to samples provided by a user.
- Inlier-based outlier detection [3], where the goal is to detect outliers from an unlabeled data set based on inlier samples.
- One-vs-rest classification [4], which is typical of problems where the negative class is too diverse and it is therefore difficult to collect and label enough negative samples.
- Open set recognition [5], where testing classes are unknown at training time and therefore the exploitation of unlabeled data may help to learn more robust concepts.

Naive approaches are proposed to address PU learning. In particular, it is possible to distinguish between solutions that firstly identify reliable negative samples from the unlabeled data based on some heuristics and secondly train a standard binary classifier, and solutions based on binary classifiers that consider all unlabeled data as negative. In the former case, the approaches are heavily dependent on heuristics, while in the latter case, the approaches are prone to the problem of wrong label assignment.

The recent works in [6] and [7] formulate PU learning as optimization problems under the framework of statistical learning theory [8]. In both works, there is theoretical analysis of the problem with the derivation of generalization error bounds and study of the optimality of the obtained solutions. Nevertheless, there is still lack of **theoretically grounded scalable** algorithms. Therefore, this work focuses on studying the problem of scalability. In particular, starting from the formulation of a convex optimization problem, we derive an efficient algorithm, that requires less storage capacity as well as faster time execution than existing state-of-the-art approaches, and prove theoretically that the

algorithm is guaranteed to converge to the exact optimal solution.

The rest of the paper is organized as follows. Section 2 reviews the related work, starting from the comparison of PU learning with one-class classification and semi-supervised learning and describing the main theoretical results achieved in PU learning. Section 3 provides a summary of the formulation of the optimization problem under the framework of statistical learning theory and enunciates for the first time the representer theorem for PU learning. Section 4 and Section 5 describe the USMO algorithm and prove its convergence, respectively. Section 6 provides a comprehensive evaluation of the proposed algorithm on a large collection of real-world data sets. Finally, we conclude this work with some considerations about future research directions.

2 LITERATURE REVIEW

PU learning is very well known in the machine learning community, since it is used in a variety of tasks ranging from matrix completion [9], multi-view learning [10], as well as semi-supervised learning [11]. It is also applied in data mining to classify data streams [12] or time series [13] and to detect events, like co-occurrences, in graphs [14].

The majority of existing works in PU learning can be classified in two broad categories, characterized by two different ways of exploiting unlabeled data. The first category consists of **two-stage** approaches [15], [16], [17], [18], which firstly extract a set of reliable negative samples from the unlabeled data and secondly use them, together with the available positive data, to train a binary classifier. These methods are mainly heuristic and their performance strongly depend on the quality of extraction of negative samples. The second category consists of single stage approaches that regard **all unlabeled data as negative samples**. Positive and negative data are then used to train different classifiers based on SVM [1], [19], neural networks [20] or kernel density estimators [21]. These approaches are subject to the problem of **wrong label assignment**, whose effect depends on the

proportion of positive samples in the unlabeled dataset. We will see later, in the discussion about theoretical studies of PU learning, how critical this issue is. For the moment, we focus on analyzing the relations of PU learning with one-class classification and semi-supervised learning, which enable us to draw some clear boundaries between these tasks and to highlight the novelty of the novelties of this work.

2.1 Comparison with one-class classification

The main goal of one-class classification (OCC) is to estimate the support of data distribution, which is extremely useful in unsupervised learning, especially in high-dimensional feature spaces, where it is very difficult to perform density estimation.

OCC is applied to many real-world problems, involving for example anomaly/novelty detection (see [22] for a recent survey and definition of anomaly detection and see [23], [24] for reviews about novelty detection). Other possible applications of OCC range from author verification in text documents [25], document retrieval [2], as well as collaborative filtering in social networks [26].

Authors in [27], [28] are among the first to develop OCC algorithms.¹ In particular, the study in [27] proposes a classifier which finds the hyperplane separating data from the origin with the maximum margin, while authors in [28] propose a classifier which finds the minimum radius hypersphere enclosing data. Despite the difference between these two approaches, it is proved in [27] that, for specific choices of kernel function (namely, in the case of translation-invariant kernels, like the Gaussian kernel), the obtained solutions are the same. Extensions of these two pioneering works, falling in the category of kernel methods, are proposed few years later. In fact, authors in [30] modify the model of [27] by incorporating a small training set of anomalies and using the centroid of this set, instead of the origin, as the reference point to find the hyperplane. Authors in [31] propose a strategy to automatically select the hyperparameters defined in [28] to increase the usability of the framework. Rather than repelling samples from a specific point, as in [27], [30], authors in [32] propose a strategy that attract samples towards the centroid. They solve a linear programming problem, where the average output of the target function computed on the training samples is minimized. Authors in [33] propose a similar strategy based on linear programming, where data are represented in a similarity/dissimilarity space. The framework is well suited for OCC applications involving strings, graphs or shapes. Other solutions different from kernel methods are also proposed. To mention a few, authors in [34] propose a neural network-based approach, where the goal is to learn the identity function. New samples are fed to the network and tested against their corresponding outputs. The test sample is considered as part of the class of interest only when the input and the output of the network are similar. Authors in [35] propose a one-class nearest neighbour, where a test sample is accepted as a member of the target class only when the distance from its neighbours is comparable to their local density. **It is worth noting that the majority of works in OCC focuses on increasing classification**

performance, rather than improving scalability. This can be arguably motivated by the fact that it is difficult in general to collect large amount of training samples from the concept/class of interest. Solutions to improve classification performance are obtained by applying classical strategies, like ensemble methods [36], bagging [37] or boosting [38] strategies. Authors in [39] argue that existing one-class classifiers fail when dealing with mixture distributions. To this purpose, they propose a multi-class classifier exploiting the supervised information of all classes of interest to refine support estimation.

A very promising solution to increase performance of OCC consists of exploiting unlabeled data, which are usually available in large quantities. As discussed in [21], standard OCC algorithms are not designed to use unlabeled data and therefore make the implicit assumption that they are uniformly distributed on the support of nominal distribution, which does not hold in general. The recent work in [40] proves that, under some simple conditions,² large amount of unlabeled data can boost performance of OCC even in comparison with completely supervised approaches. Furthermore, the exploitation of unlabeled data allows building OCC classifiers in the context of open set recognition [5], where it is essential to learn robust concepts/functions. The primary goal of PU learning is therefore to exploit this unsupervised information. **PU learning can be regarded as a generalization of OCC** [41], in the sense that it can deal and manage unlabeled data coming from more general distributions than the uniform one.

2.2 Comparison with semi-supervised learning

The idea of exploiting unlabeled data in semi-supervised learning is originally proposed by [42]. Most early studies do not provide any explanation about why the unlabeled data can be beneficial. Authors in [43] are among the first to analyze this aspect from a generative perspective. In particular, they assume that data are distributed according to a mixture of Gaussians and show that the class posterior distribution can be decomposed in two distinct terms, namely one depending on the class labels and the other one depending on the mixture components. The former can be therefore estimated using the labeled data, while the latter is estimated using the unlabeled data, thus improving the performance of the learnt classifier. Authors in [44] extend this analysis and consider that data can be correctly described by the more general class of parametric models. They show that if both the class posterior distribution and the data prior distribution are dependent on model parameters, then unlabeled examples can be exploited to learn a better set of parameters. Therefore, the key idea of semi-supervised learning is to exploit the distributional information contained in the unlabeled examples.

Many approaches have been proposed. The work in [45] provides a reference taxonomy of semi-supervised learning algorithms. In particular, we distinguish among generative approaches, like the one in [46], which exploit the unlabeled data to better estimate the class-conditional densities and infer/predict the unknown labels based on the learnt model,

1. More precisely, the term OCC was coined in 1996 [29].

2. The conditions are based on class prior and size of positive (class of interest) and negative (the rest) data.

low-density separation methods, like the work in [47], which look for decision boundaries that correctly classify labeled data and are placed in regions with few unlabeled samples (the so called low-density regions), graph-based methods, like in [48], which exploit unlabeled data to build a similarity graph and then propagate labels based on the smoothness assumption, methods based on dimensionality reduction, like in [49], which use the unlabeled samples for representation learning and then perform classification on the learnt feature representation, and disagreement-based methods, discussed in [50], which exploit the disagreement among multiple base learners to learn more robust ensemble classifiers.

The scalability issue is largely studied in the context of semi-supervised learning. For example, the work in [51] proposes a framework to solve a mixed-integer programming problem, which runs multiple times the SVM algorithm. State-of-the-art implementations of SVM (see for example LIBSVM [52]) are based mainly on decomposition methods [53], like our proposed approach. Other semi-supervised approaches look for approximations of the fitness function involved in the optimization problem, like [54], [55].

Both semi-supervised and PU learning exploit unlabeled data to learn better classifiers. Nevertheless, there are substantial differences that make semi-supervised learning not applicable to PU learning tasks. An important aspect is the fact that the majority of works in semi-supervised learning make the assumption that unlabeled data are originated only from a set of known classes (i.e. close set environment) and do not cope therefore with the presence of unknown classes in the training and test datasets (i.e. open set environment). To the best of our knowledge, only few works like the study in [56] propose semi-supervised methods capable to handle this situation. Another important and more relevant aspect is that **semi-supervised learning cannot learn a classifier when only one known class is present**, since they require at least two known classes to decide where to place the decision boundary. On the contrary, recent works in [11], [40], show that it is possible to apply PU learning algorithms to solve semi-supervised learning tasks, even in the case of open set environment.

2.3 Theoretical studies about PU learning

Inspired by the seminal work in [57] and by the first studies about OCC [27], [28], authors in [58], and later with an extended version in [59], are the first to define and theoretically analyze the problem of PU learning. In particular, the authors propose a framework based on the statistical query model [57] to have theoretical guarantees about classification performance and to derive algorithms based on decision trees. The authors study the problem of learning functions characterized by monotonic conjunctions, which are particularly useful in text classification, where documents can be represented with binary vectors, to model the presence/absence of words from an available dictionary. Instead of considering binary features, authors in [60] propose a Naive-Bayes classifier for categorical features in noisy environments. Their work is subject to the attribute independence assumption, which is useful for

estimating class-conditional densities in high-dimensional spaces, but rather limiting when compared to discriminative approaches, which directly focus on classification and avoid performing density estimation [61].

As already mentioned at the beginning of this section, the majority of PU learning studies can be classified in two main categories, viz. two-stage approaches and single stage methods that regard all unlabeled data as negative. The former are based on heuristics to select a set of reliable negative samples from the unlabeled data and are not theoretically sound, while the latter are subject to the problem of wrong label assignment. In order to understand the criticality of this issue, consider the theoretical result of consistency presented in [21].³ For any set of classifiers \mathcal{F} of Vapnik-Chervonenkis (VC) dimension V and any $\delta > 0$, there exists a constant c such that the following bounds hold with probability $1 - \delta$:

$$\begin{aligned} FNR(f) - FNR(f^*) &\leq c\epsilon_n, \\ FPR(f) - FPR(f^*) &\leq \frac{c}{1 - \pi}(\epsilon_n + \epsilon_p), \end{aligned}$$

where FPR, FNR are the false positive/negative rates, $f \in \mathcal{F}$ is the function obtained by using the above-mentioned strategy, $f^* \in \mathcal{F}$ is the optimal function having access to the ground truth, π is the positive class prior, $\epsilon_n = \sqrt{\frac{V \log(\cdot) - \log(\delta)}{n}}$ and p and n are the number of positive and "not labeled" samples, respectively. In particular, if one considers a simple scenario, where the feature space is \mathbb{R}^{100} and $V = 101$ (in the case of linear classifiers), then it is possible to learn a classifier such that, with probability of 90%, the performance do not deviate from the optimal values for more than 10% (which is equivalent to setting $\delta, \epsilon_p, \epsilon_n = 0.1$). This is guaranteed when both positive and unlabeled sets consist of at least 10^5 training samples each. This is impractical in real world applications, since collecting and labelling that amount of data is usually very expensive. The effect of wrong label assignment is even more evident for larger values of positive class prior.

Recently, authors in [6], [7] propose frameworks based on the statistical learning theory [8]. These works are free from heuristics (as opposite to two stage approaches), are not prone to the problem of wrong label assignment and are theoretically grounded, since they provide bounds on the generalization error. Nevertheless, there is lack of **theoretically grounded scalable PU learning approaches**.

3 PU LEARNING FORMULATION

Assume that we are given a training dataset $D_b = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in X, y_i \in Y\}_{i=1}^m$, where $X \subseteq \mathbb{R}^d$, $Y = \{-1, 1\}$ and each pair of samples in D_b is drawn independently from the same unknown joint distribution \mathcal{P} defined over X and Y . The goal is to learn a function f that maps the input space X into the class set Y , known as the binary classification problem. According to statistical learning theory [8], the

3. It is rewritten to be more consistent with the notation in this paper.

function f can be learnt by minimizing the risk functional \mathcal{R} , namely

$$\begin{aligned}\mathcal{R}(f) &= \sum_{y \in Y} \int \ell(f(\mathbf{x}), y) \mathcal{P}(\mathbf{x}, y) d\mathbf{x} \\ &= \pi \int \ell(f(\mathbf{x}), 1) \mathcal{P}(\mathbf{x}|y=1) d\mathbf{x} \\ &\quad + (1-\pi) \int \ell(f(\mathbf{x}), -1) \mathcal{P}(\mathbf{x}|y=-1) d\mathbf{x}\end{aligned}\quad (1)$$

where π is the positive class prior and ℓ is a loss function measuring the disagreement between the prediction and the ground truth for sample x , viz. $f(\mathbf{x})$ and y , respectively.

In PU learning, the training set is split into two parts, namely a set of samples $D_p = \{\mathbf{x}_i \in X\}_{i=1}^p$ drawn from the positive class and a set of "not labeled" samples $D_n = \{\mathbf{x}_i \in X\}_{i=1}^n$ drawn from both the positive and the negative classes. The goal is the same of the binary classification problem, but this time the supervised information is available only for one class. The learning problem can be still formulated as a risk minimization. In fact, since $\mathcal{P}(\mathbf{x}) = \pi \mathcal{P}(\mathbf{x}|y=1) + (1-\pi) \mathcal{P}(\mathbf{x}|y=-1)$, (1) can be rewritten in the following way:

$$\begin{aligned}\mathcal{R}(f) &= \pi \int \ell(f(\mathbf{x}), 1) \mathcal{P}(\mathbf{x}|y=1) d\mathbf{x} \\ &\quad + (1-\pi) \int \ell(f(\mathbf{x}), -1) \frac{\mathcal{P}(\mathbf{x}) - \pi \mathcal{P}(\mathbf{x}|y=1)}{1-\pi} d\mathbf{x} \\ &= \pi \int \tilde{\ell}(f(\mathbf{x}), 1) \mathcal{P}(\mathbf{x}|y=1) d\mathbf{x} + \int \ell(f(\mathbf{x}), -1) \mathcal{P}(\mathbf{x}) d\mathbf{x}\end{aligned}\quad (2)$$

where $\tilde{\ell}(f(\mathbf{x}), 1) = \ell(f(\mathbf{x}), 1) - \ell(f(\mathbf{x}), -1)$ is called the **composite loss** [7].

The risk functional in (2) can not be minimized since the distributions are unknown. In practice, one considers the empirical risk functional in place of (2), where expectation integrals are replaced with the empirical mean estimates computed over the available training data, namely

$$\mathcal{R}_{emp}(f) = \frac{\pi}{p} \sum_{\mathbf{x}_i \in D_p} \tilde{\ell}(f(\mathbf{x}_i), 1) + \frac{1}{n} \sum_{\mathbf{x}_i \in D_n} \ell(f(\mathbf{x}_i), -1) \quad (3)$$

The minimization of \mathcal{R}_{emp} is in general an ill-posed problem. A regularization term is usually added to \mathcal{R}_{emp} in order to restrict the solution space and to penalize complex solutions. The learning problem is therefore stated as an optimization task:

$$f^* = \arg \min_{f \in \mathcal{H}_k} \left\{ \mathcal{R}_{emp}(f) + \lambda \|f\|_{\mathcal{H}_k}^2 \right\} \quad (4)$$

where λ is a positive real parameter weighting the relative importance of the regularizer with respect to the empirical risk and $\|\cdot\|_{\mathcal{H}_k}$ is the norm associated with the function space \mathcal{H}_k . In this case, \mathcal{H}_k refers to the Reproducing Kernel Hilbert Space (RKHS) associated with its Mercer kernel $k : X \times X \rightarrow \mathbb{R}$.⁴ Considering the problem in (4), we can enunciate the representer theorem for PU learning (proof in Appendix A):

Representer Theorem 1. Given the training set $D = D_p \cup D_n$ and the Mercer kernel k associated with the RKHS \mathcal{H}_k , any minimizer $f^* \in \mathcal{H}_k$ of (4) admits the following representation

$$f^*(\mathbf{x}) = \sum_{i: \mathbf{x}_i \in D} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$$

where $\alpha_i \in \mathbb{R}$ for all i .

It is worth mentioning that many types of representer theorem have been already proposed, but none of them can be applied to the PU learning problem. Just to mention a few, authors in [63] have provided a generalized version of the classical representer theorem for classification and regression tasks, while authors in [48] have derived the representer theorem for semi-supervised learning. More recently, the study in [64] has proposed a unified view of existing representer theorems, identifying the relations between these theorems and certain classes of regularization penalties. Nevertheless, the proposed theory does not apply to problem (4) due to the hypotheses made on the empirical risk functional.

This theorem shows that it is possible to learn functions defined on possibly-infinite dimensional spaces, namely the ones induced by the kernel k , but that depend only on a finite number of parameters (namely α_i). The training focuses therefore on learning this restricted set of parameters. Another important aspect is that the representer theorem does not mention anything about the uniqueness of the solution and it only says that every minimum solution has the same parametric form. In other words, different solutions have different values of parameters. The uniqueness of the solution can be guaranteed only when the empirical risk functional in (4) is convex. A proper selection of the loss function is therefore necessary to achieve this condition. Authors in [7] have analysed the properties of loss functions for the PU learning problem and shown that a necessary condition for convexity is that the composite loss function in (3) is affine. This requirement is satisfied by some loss functions, like the squared loss, the modified Huber loss, the logistic loss, the perceptron loss and the double Hinge loss. In particular, better generalization performance can be achieved by using the double Hinge loss [7].⁵ Even, the comparison with non-convex loss functions [6], [7] suggests to use the double Hinge loss for the PU learning problem. The advantages are twofold: guarantee that the obtained solution is globally optimal, and possibility of exploiting the convex optimization theory to perform a more efficient training.

These considerations, together with the result stated by the representer theorem, allow us to rewrite problem (4) in an equivalent parametric form. In particular, by defining $\boldsymbol{\alpha} \in \mathbb{R}^{(p+n)}$ as the vector of alpha values, $\boldsymbol{\xi} \in \mathbb{R}^n$ as the vector of slack variables, $\mathbf{K} \in \mathbb{R}^{(p+n) \times (p+n)}$ as the Gram matrix computed using the training set D , and by considering without loss of generality, target functions in the form $f(\mathbf{x}) = \sum_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + \beta$, where β is the bias of f , it is possible to derive the following optimization problem (derivation in Appendix B):

$$\begin{aligned}\min_{\boldsymbol{\alpha}, \boldsymbol{\xi}, \beta} & \left\{ -c_1 \tilde{\mathbf{1}}^T \mathbf{K} \boldsymbol{\alpha} - c_1 \tilde{\mathbf{1}}^T \mathbf{1} \beta + c_2 \mathbf{1}_n^T \boldsymbol{\xi} + \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \right\} \\ \text{s.t. } & \boldsymbol{\xi} \succeq \mathbf{0}_n, \\ & \boldsymbol{\xi} \succeq \mathbf{U} \mathbf{K} \boldsymbol{\alpha} + \beta \mathbf{1}_n, \\ & \boldsymbol{\xi} \succeq \frac{1}{2} \mathbf{1}_n + \frac{1}{2} \mathbf{U} \mathbf{K} \boldsymbol{\alpha} + \frac{\beta}{2} \mathbf{1}_n,\end{aligned}\quad (5)$$

5. The double Hinge loss can be considered as the equivalent of the Hinge loss function for the binary classification problem.

4. For an overview of RKHS and their properties, see the work in [62]

where $\tilde{\mathbf{1}} = [1, \dots, 1, 0, \dots, 0]^T$ is a vector of size $p + n$ with p non-zero entries, $\mathbf{1}$ and $\mathbf{1}_n$ are unitary vectors of size $p + n$ and n , respectively, \mathbf{U} is a $n \times (p + n)$ matrix obtained through the concatenation of a $n \times p$ null matrix and an identity matrix of size n , \succeq is an element-wise operator, $c_1 = \frac{\pi}{2\lambda p}$ and $c_2 = \frac{1}{2\lambda n}$.

The equivalent dual problem of (5) is more compactly expressed as:

$$\begin{aligned} \min_{\sigma, \delta} \left\{ \frac{1}{2} \sigma^T \mathbf{U} \mathbf{K} \mathbf{U}^T \sigma - c_1 \tilde{\mathbf{1}}^T \mathbf{K} \mathbf{U}^T \sigma - \frac{1}{2} \mathbf{1}_n^T \delta \right\} \\ \text{s.t. } \mathbf{1}^T \left[c_1 \tilde{\mathbf{1}} - \mathbf{U}^T \sigma \right] = 0, \\ \sigma + \frac{1}{2} \delta \preceq c_2 \mathbf{1}_n, \\ \sigma - \frac{1}{2} \delta \succeq \mathbf{0}_n, \\ \mathbf{0}_n \preceq \delta \preceq c_2 \mathbf{1}_n, \end{aligned} \quad (6)$$

where $\sigma, \delta \in \mathbb{R}^n$ and are related to the Lagrange multipliers introduced during the derivation of the dual formulation (see Appendix B for details).

Due to linearity of constraints in (5), Slater's condition is trivially satisfied⁶ and therefore strong duality holds. This means that (6) can be solved in place of (5) to get the primal solution. The optimal α can be obtained from one of the stationarity conditions used during the Lagrangian formulation (details in Appendix B), namely using the following relation

$$\alpha = c_1 \tilde{\mathbf{1}} - \mathbf{U}^T \sigma \quad (7)$$

Note that the bias β has to be considered separately, since problem (6) does not give any information on how to compute it (this will be discussed in the next section).

It is important to point out that (6) is a quadratic programming (QP) problem that can be solved by existing numerical QP optimization libraries. Nevertheless, it is memory inefficient, since it requires to store the Gram matrix which scales quadratically with the number of training samples. Due to this problem, a modern computer cannot manage to solve (6) even for a few thousands of samples. A question therefore arises: **is it possible to efficiently find an exact solution to problem (6) without storing the whole Gram matrix?**

4 USMO ALGORITHM

In order not to store the Gram matrix, we propose an iterative algorithm that converts problem (6) into a sequence of smaller QP subproblems. Here, smaller refers to considering a subset of training samples, whose indices define the working set, and to compute and store temporarily the Gram matrix only for this reduced set. Each iteration of the USMO algorithm is composed by three main operations: the selection of the working set, called S , the computation of the Gram matrix only for samples associated to indices in S and the resolution of a QP subproblem, where only terms depending on S are considered. The details of the general algorithm are given in Algorithm 1.

It is important to mention that in principle this strategy allows decreasing the storage requirement at the expense

6. See [65] for example.

Algorithm 1 General USMO algorithm

- 1: $k \leftarrow 1$.
- 2: Initialize (σ^1, δ^1) .
- 3: **while** (σ^k, δ^k) is not a stationary point of (6) **do**
- 4: Select the working set $S \subset U = \{u : x_u \in D_n\}$ with $|S| = 2$.
- 5: Compute \mathbf{K}_{SS} , \mathbf{K}_{SP} and $\mathbf{K}_{S\bar{S}}$ where $P = \{u : x_u \in D_p\}$ and $\bar{S} = U \setminus S$.
- 6: Solve

$$\begin{aligned} \min_{\sigma_S^k, \delta_S^k} \left\{ \frac{1}{2} (\sigma_S^k)^T \mathbf{K}_{SS} \sigma_S^k + e^T \sigma_S^k - \frac{1}{2} \mathbf{1}^T \delta_S^k \right\} \\ \text{s.t. } \mathbf{1}^T \sigma_S^k = c_1 p - \mathbf{1}^T \delta_S^k \\ \sigma_S^k + \frac{1}{2} \delta_S^k \preceq c_2 \mathbf{1} \\ \sigma_S^k - \frac{1}{2} \delta_S^k \succeq \mathbf{0} \\ \mathbf{0} \preceq \delta_S^k \preceq c_2 \mathbf{1} \end{aligned} \quad (8)$$

where

$$e = \mathbf{K}_{S\bar{S}} \sigma_{\bar{S}}^k - c_1 \mathbf{K}_{SP} \mathbf{1}_p$$

and

$$\tilde{\mathbf{K}} = \begin{bmatrix} \mathbf{K}_{PP} & \mathbf{K}_{PS} & \mathbf{K}_{P\bar{S}} \\ \mathbf{K}_{SP} & \mathbf{K}_{SS} & \mathbf{K}_{S\bar{S}} \\ \mathbf{K}_{\bar{S}P} & \mathbf{K}_{\bar{S}S} & \mathbf{K}_{\bar{S}\bar{S}} \end{bmatrix}, \quad \tilde{\sigma}^k = \begin{bmatrix} \sigma_S^k \\ \sigma_{\bar{S}}^k \end{bmatrix}, \quad \tilde{\delta}^k = \begin{bmatrix} \delta_S^k \\ \delta_{\bar{S}}^k \end{bmatrix}$$

$\tilde{\mathbf{K}}$, $\tilde{\sigma}^k$ and $\tilde{\delta}^k$ are permutations of \mathbf{K} , σ^k and δ^k , respectively. In general, \mathbf{K}_{VW} is used to denote a matrix containing rows of \mathbf{K} indexed by elements in set V and columns of \mathbf{K} indexed by elements in set W .

- 7: $(\sigma_S^{k+1}, \delta_S^{k+1}) \leftarrow (\sigma_S^k, \delta_S^k)$.
 - 8: $k \leftarrow k + 1$.
 - 9: **end while**
-

of a higher computational complexity. In fact, it may happen that same samples are selected multiple times over iterations, thus requiring recomputing the same quantities, namely matrices \mathbf{K}_{SS} , \mathbf{K}_{SP} and $\mathbf{K}_{S\bar{S}}$. We will see later, how to deal with this source of inefficiency.

Another important aspect is that at each iteration only few parameters are updated (the parameters indexed by the working set S , namely σ_S^k and δ_S^k), while the others are kept fixed. Here, we consider a working set of size two, since this allows us to solve the QP subproblem (8) in an analytical way without the need of further optimization algorithms. This is discussed in the next subsection.

4.1 QP Subproblem

We start by considering the following Lemma (proof in Appendix C):

Lemma 1. Given $S = \{i, j\}$, any optimal solution $\sigma_S^* = [\sigma_i^* \ \sigma_j^*]^T$, $\delta_S^* = [\delta_i^* \ \delta_j^*]^T$ of the QP subproblem (8) has to satisfy the following condition: $\forall u : x_u \in S \wedge 0 \leq \delta_u^* \leq c_2$ either $\sigma_u^* = c_2 - \frac{\delta_u^*}{2}$ or $\sigma_u^* = \frac{\delta_u^*}{2}$.

This tells us that the optimal solution (σ_S^*, δ_S^*) assumes a specific form and therefore its computation can be per-

TABLE 1
Equations and Conditions Used to Solve the Four QP Subproblems.

Case	Equations
1	$\sigma_i^k = (a^k(k(x_i, x_i) - k(x_i, x_j)) + e_1 - e_2)/\eta$
2	$\sigma_i^k = (a^k(k(x_i, x_i) - k(x_i, x_j)) + e_1 - e_2 + 2)/\eta$
3	$\sigma_i^k = (a^k(k(x_i, x_i) - k(x_i, x_j)) + e_1 - e_2 - 2)/\eta$
4	$\sigma_i^k = (a^k(k(x_i, x_i) - k(x_i, x_j)) + e_1 - e_2)/\eta$
Case	Conditions
1	$\max\{c_2/2, a^k - c_2\} \leq \sigma_j^k \leq \min\{c_2, a^k - c_2/2\}$
2	$\max\{0, a^k - c_2\} \leq \sigma_j^k \leq \min\{c_2/2, a^k - c_2/2\}$
3	$\max\{c_2/2, a^k - c_2/2\} \leq \sigma_j^k \leq \min\{c_2, a^k\}$
4	$\max\{0, a^k - c_2/2\} \leq \sigma_j^k \leq \min\{c_2/2, a^k\}$
Note:	$a^k = c_1 p - 1/\sigma_j^k, e = [e_1, e_2]^T$ and $\eta = k(x_i, x_i) + k(x_j, x_j) - 2k(x_i, x_j)$.

formed by searching in a smaller space. In particular, four possible subspaces for search can be identified:

$$\begin{aligned}
\text{Case 1: } \sigma_i^k &= c_2 - \frac{\delta_i^k}{2} \wedge \sigma_j^k = c_2 - \frac{\delta_j^k}{2} \wedge 0 \leq \delta_i^k, \delta_j^k \leq c_2, \\
\text{Case 2: } \sigma_i^k &= c_2 - \frac{\delta_i^k}{2} \wedge \sigma_j^k = \frac{\delta_j^k}{2} \wedge 0 \leq \delta_i^k, \delta_j^k \leq c_2, \\
\text{Case 3: } \sigma_i^k &= \frac{\delta_i^k}{2} \wedge \sigma_j^k = c_2 - \frac{\delta_j^k}{2} \wedge 0 \leq \delta_i^k, \delta_j^k \leq c_2, \\
\text{Case 4: } \sigma_i^k &= \frac{\delta_i^k}{2} \wedge \sigma_j^k = \frac{\delta_j^k}{2} \wedge 0 \leq \delta_i^k, \delta_j^k \leq c_2, \quad (9)
\end{aligned}$$

Therefore, in order to solve the QP subproblem (8), one can solve four optimization problems, where the objective function is the same as (8), but the inequality constraints of (8) are simplified to (9). Indeed, each of these four subproblems can be expressed as an optimization problem of **just one variable**, by exploiting the equality constraints of both (8) and (9). Their solution can be therefore computed analytically without the need of further optimization algorithms. Table 1 reports the equations used to solve the four subproblems (we omit the derivation, which is straightforward), where σ_j^k is computed for all four cases. All other variables, namely σ_i^k , δ_i^k and δ_j^k , can be obtained in a second phase by simply exploiting the equality constraints in (8) and (9).

These equations do not guarantee in general that the inequalities in (9) are satisfied. To verify this, one can rewrite these inequalities as equivalent conditions of only σ_j^k (by exploiting the equality constraints in (8) and (9)), and check σ_j^k against them, as soon as all σ_j^k are available. In case these conditions are violated, a proper clipping is applied to σ_j^k and feasibility is therefore restored. Table 1 summarizes these checking conditions.

Finally, the minimizer of the QP subproblem (8) can be obtained by retaining only the solution with the lowest level of objective.

At each iteration, the output of the algorithm is both optimal and feasible for the QP subproblem (8). The question now is: when is it also optimal for the problem (6)? This is discussed in the next subsection.

4.2 Optimality Conditions

A problem of any optimization algorithm is to determine the stop condition. In Algorithm 1, the search of the solution is stopped as soon as some stationarity conditions are met. These conditions, called Karush-Kuhn-Tucker (KKT) conditions, represent the certificates of optimality for the obtained solution. In case of (6) they are both necessary and sufficient conditions, since the objective is convex and the constraints

are affine functions [66]. More in detail, an optimal solution has to satisfy the following relations:

$$\begin{aligned}
\frac{\partial F(\sigma, \delta)}{\partial \sigma_u} - \beta &= -\lambda_u + \mu_u, \\
\frac{\partial F(\sigma, \delta)}{\partial \delta_u} &= -\frac{\lambda_u}{2} - \frac{\mu_u}{2} - \xi_u + \eta_u, \\
\lambda_u(\sigma_u + \frac{\delta_u}{2} - c_2) &= 0, \\
\mu_u(\frac{\delta_u}{2} - \sigma_u) &= 0, \\
\xi_u(\delta_u - c_2) &= 0, \\
\eta_u \delta_u &= 0, \\
\lambda_u, \mu_u, \xi_u, \eta_u &\geq 0, \quad (10)
\end{aligned}$$

and this is valid for any component of the optimal solution, namely $\forall u : x_u \in D_n$. In (10) $F(\sigma, \delta)$ is used as an abbreviation of the objective function of (6), while $\beta, \lambda_u, \mu_u, \xi_u, \eta_u$ are the Lagrange multipliers introduced to deal with the constraints in (6). These conditions can be rewritten more compactly as:

$$\begin{aligned}
0 \leq \delta_u < c_2 \wedge \sigma_u = \frac{\delta_u}{2} &\Rightarrow \frac{\partial F(\sigma, \delta)}{\partial \sigma_u} - \beta \geq 1 \\
&\Rightarrow f(x_u) \leq -1, \\
0 \leq \delta_u < c_2 \wedge \sigma_u = c_2 - \frac{\delta_u}{2} &\Rightarrow \frac{\partial F(\sigma, \delta)}{\partial \sigma_u} - \beta \leq -1 \\
&\Rightarrow f(x_u) \geq 1, \\
\delta_u = c_2 \wedge \sigma_u = \frac{c_2}{2} &\Rightarrow -1 \leq \frac{\partial F(\sigma, \delta)}{\partial \sigma_u} - \beta \leq 1 \\
&\Rightarrow -1 \leq f(x_u) \leq 1, \quad (11)
\end{aligned}$$

In order to derive both (10) and (11), one can follow a strategy similar to the proof of Lemma 1. It is easy to verify that $\frac{\partial F(\sigma, \delta)}{\partial \sigma_u} = -f(x_u) + \beta$. Thus, (11) provides also conditions on the target function f . From now on, we will refer to (11) as the **optimality conditions**, to distinguish from approximate conditions used to deal with numerical approximations of calculators. To this aim, the **τ -optimality conditions** are introduced, namely:

$$\begin{aligned}
0 \leq \delta_u < c_2 \wedge \sigma_u = \frac{\delta_u}{2} &\Rightarrow \frac{\partial F(\sigma, \delta)}{\partial \sigma_u} - \beta \geq 1 - \frac{\tau}{2}, \\
&\Rightarrow f(x_u) \leq -1 + \frac{\tau}{2}, \\
0 \leq \delta_u < c_2 \wedge \sigma_u = c_2 - \frac{\delta_u}{2} &\Rightarrow \frac{\partial F(\sigma, \delta)}{\partial \sigma_u} - \beta \leq -1 + \frac{\tau}{2}, \\
&\Rightarrow f(x_u) \geq 1 - \frac{\tau}{2}, \\
\delta_u = c_2 \wedge \sigma_u = \frac{c_2}{2} &\Rightarrow -1 - \frac{\tau}{2} \leq \frac{\partial F(\sigma, \delta)}{\partial \sigma_u} - \beta \leq 1 + \frac{\tau}{2}, \\
&\Rightarrow -1 - \frac{\tau}{2} \leq f(x_u) \leq 1 + \frac{\tau}{2}, \quad (12)
\end{aligned}$$

where τ is a real-positive scalar used to perturb the optimality conditions.

By introducing the sets $D_1(\sigma, \delta) = \{x_u \in D_n : 0 \leq \delta_u < c_2 \wedge \sigma_u = \frac{\delta_u}{2}\}$, $D_2(\sigma, \delta) = \{x_u \in D_n : 0 \leq \delta_u < c_2 \wedge \sigma_u = c_2 - \frac{\delta_u}{2}\}$ and $D_3(\sigma, \delta) = \{x_u \in D_n : 0 < \delta_u \leq c_2 \wedge (\sigma_u = \frac{\delta_u}{2} \vee \sigma_u = c_2 - \frac{\delta_u}{2})\}$ and the quantities $m_1^{max}(\sigma, \delta) = \max_{x_u \in D_1} f(x_u)$, $m_2^{min}(\sigma, \delta) = \min_{x_u \in D_2} f(x_u)$, $m_3^{min}(\sigma, \delta) = \min_{x_u \in D_3} f(x_u)$ and $m_3^{max}(\sigma, \delta) = \max_{x_u \in D_3} f(x_u)$, called the **most critical values**, it is possible to rewrite conditions (12) in the following equivalent way:

$$\begin{aligned}
m_1^{max}(\sigma, \delta) - m_3^{min}(\sigma, \delta) &\leq \tau, \\
m_3^{max}(\sigma, \delta) - m_2^{min}(\sigma, \delta) &\leq \tau, \\
m_1^{max}(\sigma, \delta) - m_2^{min}(\sigma, \delta) + 2 &\leq \tau, \quad (13)
\end{aligned}$$

Apart from being written more compactly than (12), conditions (13) have the advantage that they can be computed without knowing the bias β . Due to the dependence on σ and δ , m_1^{max} , m_2^{min} , m_3^{min} , m_3^{max} need to be tracked and computed at each iteration in order to check τ -optimality and to decide whether to stop the algorithm.

4.3 Working Set Selection

A natural choice for selecting the working set is to look for pairs violating the τ -optimality conditions. In particular,

Definition 1. Any pair (x_i, x_j) from D_n is a **violating pair**, if and only if it satisfies the following relations:

$$\begin{aligned} x_i \in D_1, x_j \in D_3 &\Rightarrow f(x_i) - f(x_j) > \tau, \\ x_i \in D_3, x_j \in D_1 &\Rightarrow f(x_i) - f(x_j) < -\tau, \\ x_i \in D_2, x_j \in D_3 &\Rightarrow f(x_i) - f(x_j) < -\tau, \\ x_i \in D_3, x_j \in D_2 &\Rightarrow f(x_i) - f(x_j) > \tau, \\ x_i \in D_1, x_j \in D_2 &\Rightarrow f(x_i) - f(x_j) + 2 > \tau, \\ x_i \in D_2, x_j \in D_1 &\Rightarrow f(x_i) - f(x_j) - 2 < -\tau, \end{aligned} \quad (14)$$

Conditions (13) are not satisfied as long as violating pairs are found. Therefore, the algorithm keeps looking for violating pairs and use them to improve the objective function until τ -optimality is reached.

The search of violating pairs as well as the computation of the most critical values go hand in hand in the optimization and follow two different approaches. The former consists of finding violating pairs and computing the most critical values based only on a subset of samples called the **non-bound** set, namely $D_n^- = (D_1 \cap D_3) \cup (D_2 \cap D_3)$,⁷ while the latter consists of looking for violating pairs based on the whole set D_n by scanning all samples one by one. In this second approach, the most critical values are updated using the non-bound set together with the current examined sample. Only when all samples are examined, it is possible to check conditions (13), since the most critical values correspond to the original definition. The algorithm keeps using the first approach until τ -optimality for the non-bound set is reached, after that the second approach is used. This process is repeated until the τ -optimality for the whole set D_n is achieved.⁸

The motivation of having two different approaches for selecting the violating pairs and computing the most critical values is to enhance efficiency in computation. This will be clarified in the next subsection.

4.4 Function Cache and Bias Update

Recall that each iteration of the USMO algorithm is composed by three main operations, namely: the working set selection, the resolution of the associated QP subproblem, and the update of the most critical values based on the obtained solution. It is interesting to note that all these operations require to compute the target function $f(x_u)$ for all x_u belonging either to the non-bound set or to the whole set D_n , depending on the approach selected by that iteration. In fact, the stage of working set

selection requires to evaluate conditions in (14) for pairs of samples depending on f ; the equations used to solve the QP subproblem, shown in Table 1, depend on vector $e = K_{SS}\sigma_S^k - c_1 K_{SP}1_p + K_{SS}\sigma_S^{k-1} - K_{SS}\sigma_S^{k-1} = -[f(x_i) - \beta, f(x_j) - \beta]^T - K_{SS}\sigma_S^{k-1}$, which is also influenced by f ; finally, the computation of the most critical values requires to evaluate the target function f . Therefore, it is necessary to define a strategy that limits the number of times the target function is evaluated at each iteration. This can be achieved by considering the fact that the algorithm performs most of the iterations on samples in the non-bound set, while the whole set is used mainly to check if τ -optimality is reached, and then the values of the target function for those samples can be stored in a cache, called the **function cache**. Since usually $|D_n^-| \ll |D_n|$, storing $f(x_u)$ for all $x_u \in D_n^-$ is a cheap operation, which allows to save a huge amount of computation, thus increasing the computational efficiency.

At each iteration the function cache has to be updated in order to take into account the changes occurred at some of the entries of vectors σ and δ , or equivalently at some of the entries of α . By defining $\mathcal{F}^k(x_u)$ as the function cache for sample x_u at iteration k , it is possible to perform the update operation using the following relation:

$$\begin{aligned} \mathcal{F}^k(x_u) = & \mathcal{F}^{k-1}(x_u) + (\alpha_i^k - \alpha_u^{k-1})k(x_i, x_u) \\ & + (\alpha_j^k - \alpha_j^{k-1})k(x_j, x_u) \end{aligned} \quad (15)$$

Since all operations at each iteration are invariant with respect to β (because they require to evaluate differences between target function values), β can be computed at the end of the algorithm, namely when τ -optimality is reached. By exploiting the fact that the inequalities in (11) become simply equalities for samples in the non-bound set, meaning that the target function evaluated at those samples can assume only two values, 1 or -1,⁹ it is possible to compute β for each of these samples in the following way:

$$\beta_u = \begin{cases} -1 - \mathcal{F}(x_u), & \forall x_u \in D_1 \cap D_3 \\ 1 - \mathcal{F}(x_u), & \forall x_u \in D_2 \cap D_3 \end{cases} \quad (16)$$

The final β can be computed by averaging of (16) over all samples in the non-bound set, in order to reduce the effect of wrong label assignment.

4.5 Initialization

As previously mentioned, the proposed algorithm is characterized by iterations focusing either on the whole training data set or on a smaller set of non-bound samples. The formers are computationally more expensive, not only because a larger amount of samples is involved in the optimization, but also because the algorithm has to perform many evaluation operations, which can be skipped in the latter case, thanks to the exploitation of the function cache. An example of this is provided in Figure 1: the chart on the left plots the time required by the algorithm to complete the corresponding set of iterations. Peaks correspond to cases where the whole training data set is considered, while valleys represent the cases where iterations are performed over the non-bound samples. The chart on the right plots the overall objective score vs. the different sets of iterations. Jumps are associated with cases where all training samples are involved in the optimization. As soon as the algorithm

7. The term *non-bound* comes from the fact that $0 < \delta_u < c_2$ for all $x_u \in D_n^-$.

8. This can be checked only when the second approach is used and after that the whole set D_n is scanned.

9. The same principle holds for conditions (12), but in this case the inequalities are defined over arbitrary small intervals centered at 1 and -1 rather than being equalities.

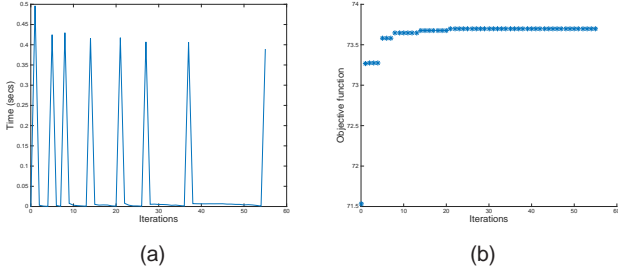


Fig. 1. (a) plot of training time over iterations, (b) learning curve expressed in terms of objective function.

approaches convergence, the contribution of the first kind of iterations becomes less and less relevant. Therefore, it is important to minimize the number of passes through the whole unlabeled data set by finding a good initialization point. To address this problem, we propose the following heuristic procedure. Labeled samples are used to train a one-class SVM [67], that is in turn used to rank the unlabeled samples according to their value of estimated target function. From this ordered list it is possible to identify groups of samples that can be associated with the cases in (11). In particular, we identify five groups of samples corresponding to the following five cases:

$$\begin{aligned}
 \delta_u^{(1)} &= 0 \wedge \sigma_u^{(1)} = 0, \\
 0 &< \delta_u^{(2)} < c_2 \wedge \sigma_u = \frac{\delta_u^{(2)}}{2}, \\
 \delta_u^{(3)} &= c_2 \wedge \sigma_u^{(3)} = \frac{c_2}{2}, \\
 0 &< \delta_u^{(4)} < c_2 \wedge \sigma_u^{(4)} = c_2 - \frac{\delta_u}{2}, \\
 \delta_u^{(5)} &= 0 \wedge \sigma_u^{(5)} = c_2,
 \end{aligned} \tag{17}$$

The size of each group as well as the initial parameters for cases in (17) can be computed by solving an optimization problem, whose objective function is defined starting from the equality constraint in (6). In particular, by defining n_1, n_2, n_3, n_4, n_5 as the sizes of the groups for the different cases and by assuming that $n_1 = (1 - \pi)an$, $n_2 = bn$, $n_3 = (1 - a - b - c)n$, $n_4 = cn$, $n_5 = \pi an$, where $a, b, c \in \mathbb{R}^+$, and that the parameters for the second and the fourth cases in (17), namely $\sigma_u^{(2)}$ and $\sigma_u^{(4)}$, are the same, the optimization problem can be formulated in the following way:

$$\begin{aligned}
 \min_{\sigma_u^{(2)}, \sigma_u^{(4)}, a, b, c} & \left\{ c_1 p - bn\sigma_u^{(2)} - cn\sigma_u^{(4)} - c_2 n \left[\pi a + \frac{1 - a - b - c}{2} \right] \right\}^2 \\
 \text{s.t. } & 0 \leq a + b + c \leq 1 - \frac{1}{n}, \\
 & \max \left\{ \frac{1}{\pi n}, \frac{1}{(1 - \pi)n} \right\} \leq a \leq \min \left\{ \frac{1}{1 - \pi}, \frac{1}{\pi} \right\}, \\
 & \frac{1}{n} \leq b, c \leq \frac{\log(n)}{n}, \\
 & 0 < \sigma_u^{(2)} < \frac{c_2}{2} \wedge \frac{c_2}{2} < \sigma_u^{(4)} < c_2,
 \end{aligned} \tag{18}$$

where the constraints in (18) can be obtained by imposing $n_1 + n_2 + n_3 + n_4 + n_5 = n$ and $1 \leq n_2, n_3, n_4, n_5 \leq n$. Furthermore, we decide to have some upper bounds for b and c to limit the size of the initial non-bound set.

In practice, after ranking the unlabeled samples through the one-class SVM and solving the optimization problem in (18), the initial solution is obtained by assigning to each sample the value of parameters corresponding to the case that sample belongs to. For example, if the samples are

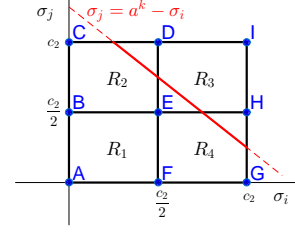


Fig. 2. Subdivision of the feasible region in the plane defined by the variables σ_i and σ_j . The red solid line represents the feasible region including the equality constraint in (8).

ranked in ascending order, then the first n_1 samples in the list have $\sigma_u = 0$ and $\delta_u = 0$, the next n_2 samples have $\sigma_u = \sigma_u^{(2)}$ and $\delta_u = 2\sigma_u$ and the others follow the same strategy.

5 THEORETICAL ANALYSIS

In this section, we present the main theoretical result, namely, we prove that Algorithm 1 converges to a τ -optimal solution.

It is important to recall that each iteration of USMO requires to solve an optimization subproblem, that depends on a single variable. In particular, if x_i and x_j correspond to the selected pair of points at one iteration, then the solution space corresponds to a line lying in the two-dimensional plane defined by the variables σ_i and σ_j . The feasible region in that plane can be subdivided into four parts, as defined according to Figure 2. These regions are considered closed sets, therefore including boundary points, like edges and corners. To consider only the interior of any set U , we use the notation $\text{int } U$. Based on these considerations, it is possible to prove the following lemma.

Lemma 2. Let the vector $z' = [\sigma'; \delta']$ be in the feasible set of (6) and (x_i, x_j) be a violating pair at point z' . Let also $z^* = [\sigma^*; \delta^*]$ be the solution obtained after applying one iteration of the Algorithm 1 using the working set $S = \{i, j\}$ and starting from z' . Then, the following results hold:

- $z^* \neq z'$,
- After the minimization step, (x_i, x_j) is no more a violating pair,
- $(\sigma_i^*, \sigma_j^*) \in \text{int } R_1 \cup \text{int } R_3 \Rightarrow f_{z^*}(x_j) - f_{z^*}(x_i) = 0$,
- $(\sigma_i^*, \sigma_j^*) \in \text{int } R_2 \Rightarrow f_{z^*}(x_j) - f_{z^*}(x_i) = 2$,
- $(\sigma_i^*, \sigma_j^*) \in \text{int } R_4 \Rightarrow f_{z^*}(x_j) - f_{z^*}(x_i) = -2$,
- $(\sigma_i^*, \sigma_j^*) \in BE \Rightarrow 0 \leq f_{z^*}(x_j) - f_{z^*}(x_i) \leq 2$,
- $(\sigma_i^*, \sigma_j^*) \in DE \Rightarrow 0 \leq f_{z^*}(x_j) - f_{z^*}(x_i) \leq 2$,
- $(\sigma_i^*, \sigma_j^*) \in EF \Rightarrow -2 \leq f_{z^*}(x_j) - f_{z^*}(x_i) \leq 0$,
- $(\sigma_i^*, \sigma_j^*) \in EH \Rightarrow -2 \leq f_{z^*}(x_j) - f_{z^*}(x_i) \leq 0$,
- $(\sigma_i^*, \sigma_j^*) \in AB \cup DI \Rightarrow f_{z^*}(x_j) - f_{z^*}(x_i) \geq 0$,
- $(\sigma_i^*, \sigma_j^*) \in AF \cup HI \Rightarrow f_{z^*}(x_j) - f_{z^*}(x_i) \leq 0$,
- $(\sigma_i^*, \sigma_j^*) \in BC \cup CD \Rightarrow f_{z^*}(x_j) - f_{z^*}(x_i) \geq 2$,
- $(\sigma_i^*, \sigma_j^*) \in FG \cup GH \Rightarrow f_{z^*}(x_j) - f_{z^*}(x_i) \leq -2$,
- $F(\sigma', \delta') - F(\sigma^*, \delta^*) > \frac{\tau}{\sqrt{2}} \|\sigma' - \sigma^*\|_2$,

where f_{z^*} represents the target function with coefficients α_i computed according to (7) using z^* .

Proof: Note that the feasible region for the QP subproblem (8) is a portion of line with negative slope lying on the plane defined by variables σ_i and σ_j (see Figure 2).

Thus, any point (σ_i, σ_j) on this line can be expressed using the following relationship:

$$\begin{aligned}\sigma_i &= \sigma'_i + t, \\ \sigma_j &= \sigma'_j - t,\end{aligned}\quad (19)$$

where $t \in \mathbb{R}$. In particular, if $t = 0$, then $(\sigma_i, \sigma_j) \equiv (\sigma'_i, \sigma'_j)$ and, if $t = t^*$, then $(\sigma_i, \sigma_j) \equiv (\sigma_i^*, \sigma_j^*)$.

Considering (19) and the fact that $\delta_{\mathbf{x}_i} = 2\sigma_i \wedge \delta_j = 2\sigma_j$ when $(\sigma_i, \sigma_j) \in R_1$, $\delta_{\mathbf{x}_i} = 2\sigma_i \wedge \delta_j = 2c_2 - 2\sigma_j$ when $(\sigma_i, \sigma_j) \in R_2$, $\delta_{\mathbf{x}_i} = c_2 - 2\sigma_i \wedge \delta_j = c_2 - 2\sigma_j$ when $(\sigma_i, \sigma_j) \in R_3$ and $\delta_{\mathbf{x}_i} = c_2 - 2\sigma_i \wedge \delta_j = 2\sigma_j$ when $(\sigma_i, \sigma_j) \in R_4$, it is possible to rewrite the objective in (8) as a function of t , namely:

$$\begin{aligned}\phi(t) &= \frac{1}{2}(\sigma'_i + t)^2 k(\mathbf{x}_i, \mathbf{x}_i) + \frac{1}{2}(\sigma'_j - t)^2 k(\mathbf{x}_j, \mathbf{x}_j) + \\ &\quad (\sigma'_i + t)(\sigma'_j - t)k(\mathbf{x}_i, \mathbf{x}_j) + h_z(t)\end{aligned}\quad (20)$$

where h_z is a function defined in the following way:

$$h_z(t) = \begin{cases} (e_1 - 1)(\sigma'_i + t) + (e_2 - 1)(\sigma'_j - t), & (\sigma_i, \sigma_j) \in R_1, \\ (e_1 - 1)(\sigma'_i + t) + (e_2 + 1)(\sigma'_j - t) - c_2, & (\sigma_i, \sigma_j) \in R_2, \\ (e_1 + 1)(\sigma'_i + t) + (e_2 + 1)(\sigma'_j - t) - 2c_2, & (\sigma_i, \sigma_j) \in R_3, \\ (e_1 + 1)(\sigma'_i + t) + (e_2 - 1)(\sigma'_j - t) - c_2, & (\sigma_i, \sigma_j) \in R_4, \end{cases}$$

Note that $\frac{d^2\phi(t)}{dt^2} = k(\mathbf{x}_i, \mathbf{x}_i) + k(\mathbf{x}_j, \mathbf{x}_j) - 2k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ (k is a Mercer kernel), meaning that (20) is convex.

If $(\sigma_i^*, \sigma_j^*) \in \text{int } R_1$, then (σ_i^*, σ_j^*) is the minimum and $\frac{d\phi(t^*)}{dt} = 0$. Since $\frac{d\phi(t^*)}{dt} = f_{\mathbf{z}^*}(\mathbf{x}_j) - f_{\mathbf{z}^*}(\mathbf{x}_i) = 0$, the first and the second conditions in (14), which are the only possibilities to have a violating pair, are not satisfied. Therefore, $(\mathbf{x}_i, \mathbf{x}_j)$ is not violating at point \mathbf{z}^* , but it is violating at \mathbf{z}' , implying that $\mathbf{z}^* \neq \mathbf{z}'$. The same situation holds for $(\sigma_i^*, \sigma_j^*) \in \text{int } R_3$ and this proves statement (c).¹⁰ Statements (d) and (e) can be proven in the same way, considering that the admissible conditions to have a violating pair are the first, the fourth and the fifth conditions for the former case and the second, the third and the sixth ones for the latter case.

If $(\sigma_i^*, \sigma_j^*) \in BE$, there are two possibilities to compute the derivative depending on the position of (σ'_i, σ'_j) , namely approaching (σ_i^*, σ_j^*) from the bottom or from the top of the constraint line. In the first case, the derivative is identified by $\frac{d\phi(t^*)}{dt^-}$, while in the second case by $\frac{d\phi(t^*)}{dt^+}$. Since (σ_i^*, σ_j^*) is the minimum and due to the convexity of function $\phi(t)$, $\frac{d\phi(t^*)}{dt^-} \geq 0$ and $\frac{d\phi(t^*)}{dt^+} \leq 0$. Furthermore, it is easy to verify that $\frac{d\phi(t^*)}{dt^-} = f_{\mathbf{z}^*}(\mathbf{x}_j) - f_{\mathbf{z}^*}(\mathbf{x}_i)$ and $\frac{d\phi(t^*)}{dt^+} = f_{\mathbf{z}^*}(\mathbf{x}_j) - f_{\mathbf{z}^*}(\mathbf{x}_i) - 2$. By combining these results, we obtain that $0 \leq f_{\mathbf{z}^*}(\mathbf{x}_j) - f_{\mathbf{z}^*}(\mathbf{x}_i) \leq 2$. This, compared with the first condition in (14), guarantees that $(\mathbf{x}_i, \mathbf{x}_j)$ is not a violating pair at \mathbf{z}^* and therefore that $\mathbf{z}^* \neq \mathbf{z}'$. The same strategy can be applied to derive statements (g)-(o).

For the sake of notation compactness, we use $\phi'(t)$ to identify both the classical and the directional derivatives of $\phi(t)$, viz. $\frac{d\phi(t)}{dt}$, $\frac{d\phi(t^*)}{dt^-}$ and $\frac{d\phi(t^*)}{dt^+}$, respectively. Therefore, it is possible to show that $\phi(t) = \phi(0) + \phi'(0)t + \frac{\phi''(0)}{2}t^2$. Furthermore, due to the convexity of $\phi(t)$, we have that

$$\begin{aligned}\phi'(0) < 0 &\Rightarrow t_q \geq t^* > 0, \\ \phi'(0) > 0 &\Rightarrow t_q \leq t^* < 0,\end{aligned}\quad (21)$$

where $t_q = -\frac{\phi'(0)}{\phi''(0)}$ is the unconstrained minimum of $\phi(t)$. From all these considerations, we can derive the following relation:

$$\phi(t^*) \leq \phi(0) + \frac{\phi'(0)}{2}t^* \quad (22)$$

In fact, if $\phi''(0) = 0$, then (22) trivially holds. If $\phi''(0) > 0$, then

$$\phi(t^*) - \phi(0) = \frac{\phi'(0)}{2}t^* \left(\frac{2t_q - t^*}{t_q} \right) \leq \frac{\phi'(0)}{2}t^* \quad (23)$$

where the last inequality of (23) is valid because $\left(\frac{2t_q - t^*}{t_q} \right) \geq 1$, by simply applying (21).

Note also that (19) can be used to derive the following result, namely:

$$\|\sigma' - \sigma^*\|_2 = |t^*|\sqrt{2} \quad (24)$$

By combining (23) and (24) and considering that conditions (14) can be compactly rewritten as $|\phi'(0)| > \tau$, we obtain that

$$\begin{aligned}\phi(0) - \phi(t^*) &\geq -\frac{\phi'(0)}{2}t^* = \frac{|\phi'(0)|}{2}|t^*| \\ &> \frac{\tau}{2}|t^*| = \frac{\tau}{2\sqrt{2}}\|\sigma' - \sigma^*\|_2,\end{aligned}\quad (25)$$

Finally, statement (p) is obtained from (25), by taking into account that $\phi(t^*) = F(\sigma^*, \delta^*)$ and $\phi(0) = F(\sigma', \delta')$. \square

Lemma 2 states that each iteration of Algorithm 1 generates a solution that is τ -optimal for the indices in the working set S .

The convergence of USMO to a τ -optimal solution can be proven by contradiction by assuming that the algorithm proceeds indefinitely. This is equivalent to assume that $(\mathbf{x}_{i^k}, \mathbf{x}_{j^k})$ is violating $\forall k \geq 0$, where (i^k, j^k) represents the pair of indices selected at iteration k .

Since $\{F(\sigma^k, \delta^k)\}$ is a decreasing sequence (due to the fact that $\mathbf{z}^k \neq \mathbf{z}^{k+1} \forall k \geq 0$ ¹¹ and that the algorithm minimizes the objective function at each iteration) and bounded below (due to the existence of an unknown global optimum), it is convergent. By exploiting this fact and by considering that $\frac{2\sqrt{2}}{\tau}[F(\sigma^k, \delta^k) - F(\sigma^{k+l}, \delta^{k+l})] > \|\sigma^k - \sigma^{k+l}\|_2, \forall k, l \geq 0$, which can be obtained from (p) of Lemma 2 by applying l times the triangle inequality, it is possible to conclude that $\{\sigma^k\}$ is a Cauchy sequence. Therefore, since the sequence lies also in a closed feasible set, it is convergent. In other words, we have that $\sigma^k \rightarrow \bar{\sigma}$ for $k \rightarrow \infty$, meaning that Algorithm 1 produces a **convergent sequence of points**. Now, it is important to understand if this sequence converges to a τ -optimal solution.

First of all, let us define the set of indices that are encountered/selected by the algorithm infinitely many times:

$$I_\infty = \{(\mu, \nu) : \exists \{k_t\} \subset \{k\}, (i^{k_t}, j^{k_t}) = (\mu, \nu), \forall t \in \mathbb{N}\} \quad (26)$$

$\{k_t\}$ is therefore a subsequence of $\{k\}$. It is also important to mention that since the number of iterations is infinite and the number of samples is finite, I_∞ cannot be an empty set. Based on this consideration, we define $\mathbf{v}_{\mu\nu}$ as the vector, whose elements are the entries at position μ and ν of a general vector \mathbf{v} , and provide the following lemma.

Lemma 3. Assume $(\mu, \nu) \in I_\infty$ and let $\{k_t\}$ be the sequence of indices for which $(i^{k_t}, j^{k_t}) = (\mu, \nu)$. Then,

10. In this case, the admissible conditions for violation are the second and the third conditions in (14).

11. Statement (a) of Lemma 2.

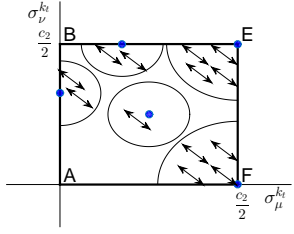


Fig. 3. Example of transitions performed by a minimization step of Algorithm 1 for different locations of $\bar{\sigma}_{\mu,\nu}$ (highlighted by blue points) and for sufficiently large number of iterations.

- (a) $\forall \epsilon > 0, \exists \hat{t} > 0: \forall t \geq \hat{t}, \|\sigma_{\mu\nu}^{k_t} - \bar{\sigma}_{\mu\nu}\| < \epsilon$ and $\|\sigma_{\mu\nu}^{k_t+1} - \bar{\sigma}_{\mu\nu}\| < \epsilon$
- (b) $f_{\sigma^{k_t}}(\mathbf{x}_\mu) - f_{\sigma^{k_t}}(\mathbf{x}_\nu) > \tau \Rightarrow f_{\bar{\sigma}}(\mathbf{x}_\mu) - f_{\bar{\sigma}}(\mathbf{x}_\nu) \geq \tau$
- (c) $f_{\sigma^{k_t}}(\mathbf{x}_\mu) - f_{\sigma^{k_t}}(\mathbf{x}_\nu) < -\tau \Rightarrow f_{\bar{\sigma}}(\mathbf{x}_\mu) - f_{\bar{\sigma}}(\mathbf{x}_\nu) \leq -\tau$
- (d) $f_{\sigma^{k_t}}(\mathbf{x}_\mu) - f_{\sigma^{k_t}}(\mathbf{x}_\nu) > \tau - 2 \Rightarrow f_{\bar{\sigma}}(\mathbf{x}_\mu) - f_{\bar{\sigma}}(\mathbf{x}_\nu) \geq \tau - 2$
- (e) $f_{\sigma^{k_t}}(\mathbf{x}_\mu) - f_{\sigma^{k_t}}(\mathbf{x}_\nu) < -\tau + 2 \Rightarrow f_{\bar{\sigma}}(\mathbf{x}_\mu) - f_{\bar{\sigma}}(\mathbf{x}_\nu) \leq -\tau + 2$

where $f_{\sigma^{k_t}}$, $f_{\bar{\sigma}}$ represent the target function with coefficients α_i computed according to (7) using σ^{k_t} and $\bar{\sigma}$, respectively.

Proof: Since $\{\sigma^k\}$ is convergent and $\{k_t\}, \{k_t + 1\}$ are subsequences of $\{k\}$, $\{\sigma^{k_t}\}$ and $\{\sigma^{k_t+1}\}$ are also convergent sequences. In other words, $\exists \hat{t} > 0$ such that $\|\sigma^{k_t} - \bar{\sigma}\| < \epsilon$ and $\|\sigma^{k_t+1} - \bar{\sigma}\| < \epsilon$. Furthermore, $\|\sigma^{k_t} - \bar{\sigma}\| \geq \|\sigma_{\mu\nu}^{k_t} - \bar{\sigma}_{\mu\nu}\|$ and $\|\sigma^{k_t+1} - \bar{\sigma}\| \geq \|\sigma_{\mu\nu}^{k_t+1} - \bar{\sigma}_{\mu\nu}\|$. By combining these two results, we obtain statement (a).

Concerning statement (b), we have that $f_{\sigma^{k_t}}(\mathbf{x}_\mu) - f_{\sigma^{k_t}}(\mathbf{x}_\nu) > \tau$. Furthermore, from convergence of $\{\sigma^{k_t}\}$ and continuity of f , we obtain that $\forall \epsilon > 0, \exists \hat{t} \geq \hat{t}: \forall t \geq \hat{t}, -\epsilon \leq f_{\sigma^{k_t}}(\mathbf{x}_\mu) - f_{\bar{\sigma}}(\mathbf{x}_\mu) \leq \epsilon$ and $-\epsilon \leq f_{\sigma^{k_t}}(\mathbf{x}_\nu) - f_{\bar{\sigma}}(\mathbf{x}_\nu) \leq \epsilon$, meaning that both $\{f_{\sigma^{k_t}}(\mathbf{x}_\mu)\}$ and $\{f_{\sigma^{k_t}}(\mathbf{x}_\nu)\}$ are convergent. Therefore, $f_{\sigma^{k_t}}(\mathbf{x}_\mu) - f_{\sigma^{k_t}}(\mathbf{x}_\nu) > \tau$ can be rewritten as

$$f_{\sigma^{k_t}}(\mathbf{x}_\mu) - f_{\sigma^{k_t}}(\mathbf{x}_\nu) + f_{\bar{\sigma}}(\mathbf{x}_\mu) - f_{\bar{\sigma}}(\mathbf{x}_\mu) + f_{\bar{\sigma}}(\mathbf{x}_\nu) - f_{\bar{\sigma}}(\mathbf{x}_\nu) > \tau$$

and by applying the information about the convergence of both $\{f_{\sigma^{k_t}}(\mathbf{x}_\mu)\}$ and $\{f_{\sigma^{k_t}}(\mathbf{x}_\nu)\}$, we get that

$$f_{\bar{\sigma}}(\mathbf{x}_\mu) - f_{\bar{\sigma}}(\mathbf{x}_\nu) > \tau - 2\epsilon$$

which is valid $\forall \epsilon > 0$ and therefore proves statement (b). All other statements, namely (c)-(e), can be proven using the same approach. \square

Lemma 3 states some conditions about the final target function and also states that the sequence output by Algorithm 1, after a sufficiently large number of iterations, is enclosed in a ball centered at $\bar{\sigma}$. This aspect is shown in Figure 3 for R_1 and for different possible locations of $\bar{\sigma}_{\mu,\nu}$. The same picture shows also the possible transitions that may happen at each iteration. In particular, we see that for $\bar{\sigma}_{\mu,\nu}$ lying on corners and edges, different kinds of transitions exist. In fact, we find transitions from border to inner points and viceversa, and transitions from inner points to inner points. These are identified as $bd \rightarrow bd$, $bd \rightarrow int$, $int \rightarrow bd$ and $int \rightarrow int$, respectively. Note that for $\bar{\sigma}_{\mu,\nu}$ not lying on borders, $int \rightarrow int$ is the only available kind of transition. Based on these considerations, it is possible to prove the following lemma.

Lemma 4. Let (μ, ν) , $\{k_t\}$, \hat{t} and ϵ be defined according to Lemma 3. Then, $\exists \hat{t} \geq \hat{t}$ such that $\forall t \geq \hat{t}$ and for sequence

$\{k_t\}$ the only allowed transitions are $int \rightarrow bd$ and $bd \rightarrow bd$.

Proof: Consider region R_1 and $(\bar{\sigma}_\mu, \bar{\sigma}_\nu) \in int R_1$. Then, the only admissible type of transitions for this case is $int \rightarrow int$. Therefore, based on statement (c) of Lemma 2 (and thanks also to statement (a) of Lemma 3), we obtain that $\forall t \geq \hat{t}, f_{\sigma^{k_t+1}}(\mathbf{x}_\mu) - f_{\sigma^{k_t+1}}(\mathbf{x}_\nu) = 0$. By exploiting this fact, the continuity of f and the convergence of $\{\sigma^{k_t+1}\}$, it is possible to show that

$$f_{\bar{\sigma}}(\mathbf{x}_\mu) - f_{\bar{\sigma}}(\mathbf{x}_\nu) = 0 \quad (27)$$

Furthermore, since $(\mathbf{x}_\mu, \mathbf{x}_\nu)$ is a violating pair at all iterations and $\forall t \geq \hat{t}, \sigma_{\mu\nu}^{k_t} \in int R_1$ (due to statement (a) of Lemma 3), $(\mathbf{x}_\mu, \mathbf{x}_\nu)$ has to satisfy conditions (b) or (c) of Lemma 3. These conditions are in contradiction with (27), meaning that the $int \rightarrow int$ transition is not allowed in this case.

Consider now region R_1 and $(\bar{\sigma}_\mu, \bar{\sigma}_\nu) \in E$, or equivalently $(\bar{\sigma}_\mu, \bar{\sigma}_\nu) \in A$. This time, the potential transitions are $bd \rightarrow bd$, $bd \rightarrow int$, $int \rightarrow bd$ and $int \rightarrow int$. Nevertheless, it is always possible to define a subsequence containing only either $int \rightarrow int$ or $bd \rightarrow int$ and obtain therefore conclusions similar to the previous case. In fact, both $int \rightarrow int$ and $bd \rightarrow int$ are not allowed transitions.

The same results can be obtained in a similar way for other edges, corners of R_1 as well as for points in R_3 , upon selection of the proper conditions in Lemma 2.

Consider now region R_2 and $(\bar{\sigma}_\mu, \bar{\sigma}_\nu) \in int R_2$. The only admissible transition in this case is $int \rightarrow int$. From statement (d) of Lemma 2, we have that $\forall t \geq \hat{t}, f_{\sigma^{k_t+1}}(\mathbf{x}_\mu) - f_{\sigma^{k_t+1}}(\mathbf{x}_\nu) = -2$ and, from the continuity of f and the convergence of $\{\sigma^{k_t+1}\}$, it is possible to show that

$$f_{\bar{\sigma}}(\mathbf{x}_\mu) - f_{\bar{\sigma}}(\mathbf{x}_\nu) = -2 \quad (28)$$

Furthermore, since $(\mathbf{x}_\mu, \mathbf{x}_\nu)$ is a violating pair at all iterations and $\forall t \geq \hat{t}, \sigma_{\mu\nu}^{k_t} \in int R_2$, $(\mathbf{x}_\mu, \mathbf{x}_\nu)$ has to satisfy conditions (b) or (d) of Lemma 3. These conditions are in contradiction with (28), meaning that the $int \rightarrow int$ transition is not valid.

For all corners and edges of R_2 , as well as for all points in R_4 , it is possible to show that $int \rightarrow int$ and $bd \rightarrow int$ are not valid transitions. The proof is similar to the previous cases. Therefore, the only admissible transitions after a sufficiently large number of iterations are $int \rightarrow bd$ and $bd \rightarrow bd$. \square

It is interesting to note that each transition $int \rightarrow bd$ increases the number of components of σ belonging to borders of the four regions, by one or two, while each transition $bd \rightarrow bd$ leaves it unchanged. Since this number is bounded by n , transition $int \rightarrow bd$ cannot appear infinitely many times. Therefore, $\exists t^* \geq \hat{t}, \forall t \geq t^*, bd \rightarrow bd$ is the only valid transition.

Note that $bd \rightarrow bd$ may happen only when $(\bar{\sigma}_\mu, \bar{\sigma}_\nu)$ is located at some specific corners of the feasible region, namely corners A or E for region R_1 , corners B or C for region R_2 , corners E or I for region R_3 and corners F or H for region R_4 . For all cases, it is possible to define a subsequence that goes only from a vertical to a horizontal border and a subsequence that goes only from a horizontal to a vertical border. Without loss of generality, we can consider a specific case, namely $(\bar{\sigma}_\mu, \bar{\sigma}_\nu) \in A$. Note that for the first subsequence, $f_{\sigma^{k_t+1}}(\mathbf{x}_\mu) - f_{\sigma^{k_t+1}}(\mathbf{x}_\nu) < -\tau$, since $(\mathbf{x}_\mu, \mathbf{x}_\nu)$ has to be a violating pair in order not to stop the iterations and therefore,

from statement (c) of Lemma 3, $f_{\sigma}(\mathbf{x}_{\mu}) - f_{\sigma}(\mathbf{x}_{\nu}) < -\tau$. For the second subsequence, $f_{\sigma^{k_t+1}}(\mathbf{x}_{\mu}) - f_{\sigma^{k_t+1}}(\mathbf{x}_{\nu}) > \tau$ and consequently $f_{\sigma}(\mathbf{x}_{\mu}) - f_{\sigma}(\mathbf{x}_{\nu}) > \tau$. This leads to a contradiction which holds $\forall(\mu, \nu) \in I_{\infty}$. Therefore, the assumption that Algorithm 1 proceeds indefinitely is not verified. In other words, there exists an iteration at which the algorithm stops because a τ -optimal solution is obtained.

6 EXPERIMENTAL RESULTS

TABLE 2
Characteristics of data sets.

Data	# Instances	# Features
<i>Australian</i>	690	42
<i>Clean1</i>	476	166
<i>Diabetes</i>	768	8
<i>Heart</i>	270	9
<i>Heart-statlog</i>	270	13
<i>House</i>	232	16
<i>House-votes</i>	435	16
<i>Ionosphere</i>	351	33
<i>Isolet</i>	600	51
<i>Krvskp</i>	3196	36
<i>Liverdisorders</i>	345	6
<i>Spectf</i>	349	44
<i>Bank-marketing</i>	28000	20
<i>Adult</i>	32562	123
<i>Statlog (shuttle)</i>	43500	9
<i>Mnist</i>	60000	784
<i>Poker-hand</i>	1000000	10

In this section, comprehensive evaluations are performed to verify the effectiveness of USMO. The proposed algorithm is compared against [7] and [6]. The code of USMO is developed in MATLAB and it is available at XXXX XXXX.¹² Competitors are also implemented in MATLAB to guarantee fair comparisons. In particular, the method in [7] solves problem (6) using the MATLAB built-in function *quadprog*, combined with the second-order primal-dual interior point algorithm [68], while the method in [6] solves problem (4) with the ramp loss function using the *quadprog* function combined with the concave-convex procedure [69]. Experiments are run on a machine with 16 2.40 GHz cores and 64GB RAM. A large collection of real-world data sets from the UCI repository is used, namely 17 data sets, 12 of which contain few hundreds/thousands of samples, while the remaining 5 are consistently bigger. Table 2 shows some of their statistics.

Since both USMO and [7] solve the same optimization problem, we firstly investigate **whether the two algorithms achieve same solutions**. To do this, we consider the F-measure in a transductive setting, to access the generalization performance, on all small-scale data sets and under different configurations of hyperparameters and kernel function used. In particular, we consider different values of λ , viz. 0.0001, 0.001, 0.01, 0.1, and use the linear and the Gaussian kernels.¹³ In all experiments, only 20% of positive samples are labeled, while the remaining ones are considered unlabeled. Tables 3-6 show the results using the linear and the Gaussian kernel, respectively. In the majority

of cases, both algorithms achieve identical performance, thus same solutions. This fact is a direct consequence of the theory proved in Section 5, since USMO is guaranteed to converge to the same value of objective function obtained in [7]. For few cases, the small differences in performance are due to numerical approximations involved during computation. Concerning the training time, USMO outperforms the competitor in almost all cases when using the linear kernel, while obtains similar performance when using the Gaussian kernel. In the former case, the advantage of USMO is twofold, namely linear storage complexity, instead of quadratic, and faster convergence rate, while in the latter case, there is a clear advantage only in terms of storage complexity. In all cases **USMO is therefore able to achieve the same solution of [7] in a more efficient way**.

Secondly, we investigate **the quality of solutions obtained by USMO and [6]**. To do this, we adopt the same configurations used in previous analysis. Tables 3-6 show the results using the linear and the Gaussian kernel, respectively. On average, USMO is performing better than [6]. There are cases in which the two methods achieve very different performance, for example see the *house-votes*, the *ionosphere*, the *liverdisorders* and the *spectf* datasets on Table 5. This is mainly due to the fact that the optimization problem solved in [6] is not convex, meaning that multiple local solutions are available. The choice of the starting point becomes therefore a critical operation, since it directly impacts the quality of the obtained solution. Furthermore, it is important to mention that USMO has a linear storage complexity instead of quadratic, as for [6].

Finally, we investigate **the performance of all algorithms on problems of larger scale**. We use the 5 larger data sets and test the methods with a fixed number of positive labeled samples, viz. 100 instances randomly sampled from the positive class, and an increasing number of unlabeled samples. Regarding the multi-class data sets, namely *statlog*, *MNIST* and *poker-hand*, each of the available classes is considered positive and compared against the others, thus producing different experiments. In all cases, the linear kernel is used. Figure 4 and Figure 5 show all training time curves as well as the learning trends for the generalization performance (estimated over the test sets and using the F-measure). It is evident that the advantage of USMO over [7] and [6], in terms of both speed and storage, increases with the number of unlabeled samples. When the number of unlabeled samples exceeds 10000 units, the method in [7] and [6] becomes computationally intractable (in terms of storage) and only USMO can still provide a solution. Furthermore, USMO outperforms other approaches in terms of generalization performance on almost all cases.

7 CONCLUSION

In this work an efficient algorithm for PU learning is proposed. Theoretical analysis is provided to ensure that the obtained solution recovers the optimum of the objective function. Experimental evaluation assesses the efficiency of the proposed method showing its potential applicability to real-world problems involving PU learning. Future research will extend this work and consider one-shot learning as well as representation learning.

12. Except for the initialization, which uses LIBSVM [52] to run the one-class classifier.

13. The positive class prior π is set to the class proportion in the training data sets. Methods like [21], [70], [71] can be used to estimate it.

APPENDIX A

PROOF OF THE REPRESENTER THEOREM

Similarly to [63], define Φ as the set, whose elements are the representers of the training dataset $D = D_p \cup D_n$, namely $\Psi = \{\varphi_{\mathbf{x}_i} \in \mathcal{H}_k | i : \mathbf{x}_i \in D\}$. Be \mathcal{H}_Ψ the linear subspace of \mathcal{H}_k spanned by the elements in Ψ and $\bar{\mathcal{H}}_\Psi$ its orthogonal complement, such that $\mathcal{H}_k = \mathcal{H}_\Psi \oplus \bar{\mathcal{H}}_\Psi$:

$$\begin{aligned}\mathcal{H}_\Psi &= \left\{ g \in \mathcal{H}_k | g = \sum_{i:\mathbf{x}_i \in D} \alpha_i \varphi_{\mathbf{x}_i}, \alpha_i \in \mathbb{R} \right\} \\ \bar{\mathcal{H}}_\Psi &= \{ h \in \mathcal{H}_k | \langle h, g \rangle_{\mathcal{H}_k} = 0, \forall g \in \mathcal{H}_\Psi \}\end{aligned}$$

Therefore, any function $f \in \mathcal{H}_k$ can be decomposed in two orthogonal components, namely $f = f^* + f^\perp$ where $f^* \in \mathcal{H}_\Psi$ and $f^\perp \in \bar{\mathcal{H}}_\Psi$. Evaluating the function f at the training point \mathbf{x}_j is performed by exploiting the previous properties, viz.

$$\begin{aligned}f(\mathbf{x}_j) &= \langle \varphi_{\mathbf{x}_j}, f^* + f^\perp \rangle_{\mathcal{H}_k} \\ &= \langle \varphi_{\mathbf{x}_j}, f^* \rangle_{\mathcal{H}_k} \\ &= \langle \varphi_{\mathbf{x}_j}, \sum_{i:\mathbf{x}_i \in D} \alpha_i \varphi_{\mathbf{x}_i} \rangle_{\mathcal{H}_k} \\ &= \sum_{i:\mathbf{x}_i \in D} \alpha_i \langle \varphi_{\mathbf{x}_j}, \varphi_{\mathbf{x}_i} \rangle_{\mathcal{H}_k} \\ &= \sum_{i:\mathbf{x}_i \in D} \alpha_i k(\mathbf{x}_j, \mathbf{x}_i) \\ &= \sum_{i:\mathbf{x}_i \in D} \alpha_i k(\mathbf{x}_i, \mathbf{x}_j)\end{aligned}$$

where the first and second equalities are due to the reproducing property of RKHS and orthogonality, respectively. The third and the fourth equalities are simple application of the inner product properties. The fifth equality holds by the definition of reproducing kernel, while the last one is valid thanks to the symmetry of any Mercer kernel. This relation highlights the fact that the evaluation of any function f at any training point \mathbf{x}_j is independent of f^\perp . Consequently, since $R_{emp}(f)$ is a functional of f evaluated at all samples of the training dataset, we have that $R_{emp}(f)$ is also independent of f^\perp . In other words, $R_{emp}(f) = R_{emp}(f^*)$. Furthermore, thanks to the orthogonality property, one can express the regularization term in (4) in the following way:

$$\|f\|_{\mathcal{H}_k}^2 = \|f^* + f^\perp\|_{\mathcal{H}_k}^2 = \|f^*\|_{\mathcal{H}_k}^2 + \|f^\perp\|_{\mathcal{H}_k}^2$$

The objective function in (4) can be therefore lower bounded in the following way:

$$\begin{aligned}\mathcal{R}_{emp}(f) + \lambda \|f\|_{\mathcal{H}_k}^2 &= \mathcal{R}_{emp}(f^*) + \lambda \|f^*\|_{\mathcal{H}_k}^2 + \lambda \|f^\perp\|_{\mathcal{H}_k}^2 \\ &\geq \mathcal{R}_{emp}(f^*) + \lambda \|f^*\|_{\mathcal{H}_k}^2\end{aligned}$$

which is valid for any $f \in \mathcal{H}_k$. f^* is therefore the minimizer of (4) and it assumes the following form:

$$f^*(\mathbf{x}) = \sum_{i:\mathbf{x}_i \in D} \alpha_i \langle \varphi_{\mathbf{x}_i}, \varphi_{\mathbf{x}} \rangle_{\mathcal{H}_k} = \sum_{i:\mathbf{x}_i \in D} \alpha_i k(\mathbf{x}_i, \mathbf{x})$$

Q.E.D.

APPENDIX B

PU LEARNING FORMULATION

B.1 Derivation of the primal problem

By taking into account the definition of the double Hinge loss function and its composite loss, namely $\ell(x, y) = \max\{-xy, \max\{0, \frac{1}{2} - \frac{xy}{2}\}\}$ and $\tilde{\ell}(x, y) = -xy$, we can

express the empirical risk functional (3) in the following way:

$$-\frac{\pi}{p} \sum_{i:\mathbf{x}_i \in D_p} f(\mathbf{x}_i) + \frac{1}{n} \sum_{i:\mathbf{x}_i \in D_n} \max \left\{ f(\mathbf{x}_i), \max \left\{ 0, \frac{1}{2} + \frac{f(\mathbf{x}_i)}{2} \right\} \right\} \quad (29)$$

and by exploiting the result stated by the representer theorem, the optimization problem (4) becomes:

$$\begin{aligned}\min_{\alpha, \xi, \beta} & \left\{ -c_1 \sum_{i:\mathbf{x}_i \in D_p} \left(\sum_{j:\mathbf{x}_j \in D} \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + \beta \right) + c_2 \sum_{i:\mathbf{x}_i \in D_n} \xi_i \right. \\ & \left. + \frac{1}{2} \sum_{i:\mathbf{x}_i \in D} \sum_{j:\mathbf{x}_j \in D} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \right\} \\ \text{s.t. } & \xi_i \geq 0, \\ & \xi_i \geq \sum_{j:\mathbf{x}_j \in D} \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + \beta, \\ & \xi_i \geq \frac{1}{2} + \frac{1}{2} \left(\sum_{j:\mathbf{x}_j \in D} \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + \beta \right), \quad (30)\end{aligned}$$

where $c_1 = \frac{\pi}{2\lambda p}$, $c_2 = \frac{1}{2\lambda n}$ and ξ_i is the slack variable associated with sample $\mathbf{x}_i \in D_n$. Notice that slack variables are used to make the objective function differentiable.

Finally, by using vector notation, (30) can be rewritten in a more compact form:

$$\begin{aligned}\min_{\alpha, \xi, \beta} & \left\{ -c_1 \tilde{\mathbf{1}}^T \mathbf{K} \alpha - c_1 \tilde{\mathbf{1}}^T \mathbf{1} \beta + c_2 \mathbf{1}_n^T \xi + \frac{1}{2} \alpha^T \mathbf{K} \alpha \right\} \\ \text{s.t. } & \xi \succeq \mathbf{0}_n, \\ & \xi \succeq \mathbf{U} \mathbf{K} \alpha + \beta \mathbf{1}_n, \\ & \xi \succeq \frac{1}{2} \mathbf{1}_n + \frac{1}{2} \mathbf{U} \mathbf{K} \alpha + \frac{\beta}{2} \mathbf{1}_n,\end{aligned}$$

B.2 Derivation of the dual problem

The Lagrangian function for problem (5) is defined as follows:

$$\begin{aligned}\mathcal{L}(\alpha, \xi, \beta, \gamma, \delta) &= \frac{1}{2} \alpha^T \mathbf{K} \alpha - c_1 \tilde{\mathbf{1}}^T \mathbf{K} \alpha - c_1 \tilde{\mathbf{1}}^T \mathbf{1} \beta + c_2 \mathbf{1}_n^T \xi \\ &\quad - \eta^T \xi + \gamma^T (\mathbf{U} \mathbf{K} \alpha + \beta \mathbf{1}_n - \xi) \\ &\quad + \delta^T \left(\frac{1}{2} \mathbf{1}_n + \frac{1}{2} \mathbf{U} \mathbf{K} \alpha + \frac{\beta}{2} \mathbf{1}_n - \xi \right)\end{aligned}$$

where $\eta \succeq \mathbf{0}_n$, $\gamma \succeq \mathbf{0}_n$ and $\delta \succeq \mathbf{0}_n$ are vectors of size u containing the Lagrange multipliers associated to the constraints of the primal problem. By taking the derivatives of \mathcal{L} with respect to α, ξ, β and equating them to zero, we obtain the following relations:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \alpha} = \mathbf{0}_n & \Rightarrow \alpha = c_1 \tilde{\mathbf{1}} - \mathbf{U}^T \gamma - \frac{1}{2} \mathbf{U}^T \delta, \\ \frac{\partial \mathcal{L}}{\partial \xi} = \mathbf{0} & \Rightarrow c_1 \tilde{\mathbf{1}}^T \mathbf{1} - \gamma^T \mathbf{1}_n - \frac{1}{2} \delta^T \mathbf{1}_n = 0 \\ & \Rightarrow \mathbf{1}^T \left(c_1 \tilde{\mathbf{1}} - \mathbf{U}^T \gamma - \frac{1}{2} \mathbf{U}^T \delta \right) = 0, \\ \frac{\partial \mathcal{L}}{\partial \beta} = \mathbf{0}_n & \Rightarrow c_2 \mathbf{1}_n - \eta - \gamma - \delta = \mathbf{0}_n \wedge \eta, \gamma, \delta \succeq \mathbf{0}_n \\ & \Rightarrow \gamma + \delta \preceq c_2 \mathbf{1}_n \wedge \mathbf{0}_n \preceq \gamma, \delta \preceq c_2 \mathbf{1}_n,\end{aligned}$$

which are then used to build the Lagrange dual function and consequently derive the following Lagrange dual problem (we skip the derivation due to lack of space):

$$\begin{aligned} \max_{\gamma, \delta} & \left\{ -\frac{1}{2} \left(\gamma + \frac{1}{2} \delta \right)^T \mathbf{U} \mathbf{K} \mathbf{U}^T \left(\gamma + \frac{1}{2} \delta \right) + c_1 \tilde{\mathbf{1}}^T \mathbf{K} \mathbf{U}^T \left(\gamma + \frac{1}{2} \delta \right) \right. \\ & \left. + \frac{1}{2} \mathbf{1}_n^T \delta \right\} \\ \text{s.t. } & \mathbf{1}^T \left[c_1 \tilde{\mathbf{1}} - \mathbf{U}^T \left(\gamma + \frac{1}{2} \delta \right) \right] = 0, \\ & \gamma + \delta \preceq c_2 \mathbf{1}_n, \\ & \mathbf{0}_n \preceq \gamma, \delta \preceq c_2 \mathbf{1}_n, \end{aligned}$$

(6) can be finally derived by defining $\sigma = \gamma + \frac{1}{2} \delta$ and rewriting it as a minimization problem.

APPENDIX C

PROOF OF LEMMA 1

By introducing the following notation, namely

$$\sigma_S^k = \begin{bmatrix} \sigma_i^k \\ \sigma_j^k \end{bmatrix}, \delta_S^k = \begin{bmatrix} \delta_i^k \\ \delta_j^k \end{bmatrix}, \mathbf{K}_{SS} = \begin{bmatrix} k(\mathbf{x}_i, \mathbf{x}_i) & k(\mathbf{x}_i, \mathbf{x}_j) \\ k(\mathbf{x}_j, \mathbf{x}_i) & k(\mathbf{x}_j, \mathbf{x}_j) \end{bmatrix}, \mathbf{e} = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}$$

the objective function of the QP subproblem (8) evaluated at (σ_S^k, δ_S^k) can be rewritten as:

$$F(\sigma_i^k, \sigma_j^k) - \frac{\delta_i^k}{2} - \frac{\delta_j^k}{2} \quad (31)$$

where

$$\begin{aligned} F(\sigma_i^k, \sigma_j^k) &= \frac{1}{2} \begin{bmatrix} \sigma_i^k & \sigma_j^k \end{bmatrix} \begin{bmatrix} k(\mathbf{x}_i, \mathbf{x}_i) & k(\mathbf{x}_i, \mathbf{x}_j) \\ k(\mathbf{x}_j, \mathbf{x}_i) & k(\mathbf{x}_j, \mathbf{x}_j) \end{bmatrix} \begin{bmatrix} \sigma_i^k \\ \sigma_j^k \end{bmatrix} \\ &+ \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{bmatrix} \begin{bmatrix} \sigma_i^k \\ \sigma_j^k \end{bmatrix} \end{aligned}$$

and the constraints of (8) at (σ_S^k, δ_S^k) are therefore:

$$\begin{aligned} \sigma_i^k + \sigma_j^k &= a^k, \\ \sigma_i^k + \frac{\delta_i^k}{2} &\leq c_2 \wedge \sigma_j^k + \frac{\delta_j^k}{2} \leq c_2, \\ \sigma_i^k - \frac{\delta_i^k}{2} &\geq 0 \wedge \sigma_j^k - \frac{\delta_j^k}{2} \geq 0, \\ 0 &\leq \delta_i^k, \delta_j^k \leq c_2, \end{aligned} \quad (32)$$

where $a^k = c_1 p - \mathbf{1}^T \sigma_S^k$ is a constant scalar for iteration k .

Since (σ_S^*, δ_S^*) is an optimal solution of (8), it has to satisfy the Karush-Kuhn-Tucker (KKT) conditions. In particular, the stationarity conditions can be expressed in the following way:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \sigma_i^k} &= \frac{\partial F(\sigma_i^*, \sigma_j^*)}{\sigma_i^k} + \beta + \lambda_i - \mu_i = 0, \\ \frac{\partial \mathcal{L}}{\partial \sigma_j^k} &= \frac{\partial F(\sigma_i^*, \sigma_j^*)}{\sigma_j^k} + \beta + \lambda_j - \mu_j = 0, \\ \frac{\partial \mathcal{L}}{\partial \delta_i^k} &= -\frac{1}{2} + \frac{\lambda_i}{2} + \frac{\mu_i}{2} + \xi_i - \eta_i = 0, \\ \frac{\partial \mathcal{L}}{\partial \delta_j^k} &= -\frac{1}{2} + \frac{\lambda_j}{2} + \frac{\mu_j}{2} + \xi_j - \eta_j = 0, \end{aligned} \quad (33)$$

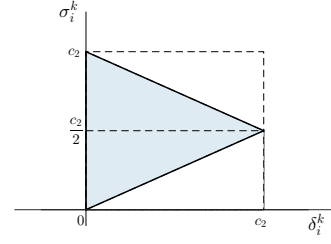


Fig. 6. Inequality constraints of the QP subproblem (8) for a given sample \mathbf{x}_i . The coloured area corresponds to the feasible region.

where \mathcal{L} is the Lagrange dual function obtained from the QP subproblem and $\beta, \lambda_i, \lambda_j, \mu_i, \mu_j, \xi_i, \xi_j, \eta_i, \eta_j$ are its Lagrange multipliers, namely:

$$\begin{aligned} \mathcal{L} &= F(\sigma_i^*, \sigma_j^*) - \frac{\delta_i^*}{2} - \frac{\delta_j^*}{2} + \beta(\sigma_i^* + \sigma_j^* - a^k) \\ &+ \lambda_i(\sigma_i^* + \frac{\delta_i^*}{2} - c_2) + \lambda_j(\sigma_j^* + \frac{\delta_j^*}{2} - c_2) \\ &- \mu_i(\sigma_i^* - \frac{\delta_i^*}{2}) - \mu_j(\sigma_j^* - \frac{\delta_j^*}{2}) \\ &+ \xi_i(\delta_i^* - c_2) + \xi_j(\delta_j^* - c_2) - \eta_i \delta_i^* - \eta_j \delta_j^* \end{aligned}$$

Now, focus on terms associated with sample \mathbf{x}_i and specifically on its inequality constraints in (32). Based on them, it is possible to distinguish the following four cases (Figure 6 helps to understand this):

$$0 \leq \delta_i^k < c_2 \wedge \frac{\delta_i^k}{2} < \sigma_i^k < c_2 - \frac{\delta_i^k}{2}, \quad (34)$$

$$0 \leq \delta_i^k < c_2 \wedge \sigma_i^k = c_2 - \frac{\delta_i^k}{2}, \quad (35)$$

$$0 \leq \delta_i^k < c_2 \wedge \sigma_i^k = \frac{\delta_i^k}{2}, \quad (36)$$

$$\delta_i^k = c_2 \wedge \sigma_i^k = \frac{\delta_i^k}{2}, \quad (37)$$

By considering the KKT complementary slackness conditions together with (33), we can derive the following statements:

$$\begin{aligned} \text{Case (34)} &\Rightarrow \lambda_i = 0, \mu_i = 0, \xi_i = 0, \eta_i \geq 0 \\ &\Rightarrow \frac{\partial F(\sigma_i^*, \sigma_j^*)}{\sigma_i^k} + \beta = 0 \wedge \eta_i = -\frac{1}{2}, \end{aligned}$$

$$\begin{aligned} \text{Case (35)} &\Rightarrow \lambda_i \geq 0, \mu_i = 0, \xi_i = 0, \eta_i \geq 0 \\ &\Rightarrow \frac{\partial F(\sigma_i^*, \sigma_j^*)}{\sigma_i^k} + \beta \leq -1 \wedge \lambda_i \geq 1 \wedge \eta_i \geq 0, \end{aligned}$$

$$\begin{aligned} \text{Case (36)} &\Rightarrow \lambda_i = 0, \mu_i \geq 0, \xi_i = 0, \eta_i \geq 0 \\ &\Rightarrow \frac{\partial F(\sigma_i^*, \sigma_j^*)}{\sigma_i^k} + \beta \geq 1 \wedge \mu_i \geq 1 \wedge \eta_i \geq 0, \end{aligned}$$

$$\begin{aligned} \text{Case (37)} &\Rightarrow \lambda_i \geq 0, \mu_i \geq 0, \xi_i \geq 0, \eta_i = 0, \\ &\Rightarrow -1 \leq \frac{\partial F(\sigma_i^*, \sigma_j^*)}{\sigma_i^k} + \beta \leq 1 \wedge -1 \leq \lambda_i, \mu_i \leq 1, \end{aligned} \quad (38)$$

The first statement in (38) is clearly a contradiction, implying that condition (34) is not valid for KKT. In other words, any optimal solution (σ_i^*, δ_i^*) does not satisfy condition (34), but only conditions (35)-(37). This fact is valid $\forall \mathbf{x}_u \in S$ due to the symmetry of the QP subproblem (8), which concludes the proof.

ACKNOWLEDGMENTS

This research was partially supported by NSFC (61333014) and the Collaborative Innovation Center of Novel Software Technology and Industrialization. We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan X Pascal machine to support this research.

REFERENCES

- [1] C. Elkan and K. Noto, "Learning Classifiers from Only Positive and Unlabeled Data," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2008, pp. 213–220.
- [2] T. Onoda, H. Murata, and S. Yamada, "One Class Support Vector Machine Based Non-Relevance Feedback Document Retrieval," in *IEEE International Joint Conference on Neural Networks*, vol. 1. IEEE, 2005, pp. 552–557.
- [3] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, "Inlier-Based Outlier Detection via Direct Density Ratio Estimation," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 223–232.
- [4] W. Li, Q. Guo, and C. Elkan, "A Positive and Unlabeled Learning Algorithm for One-Class Classification of Remote-Sensing Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 2, pp. 717–725, 2011.
- [5] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward Open Set Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2013.
- [6] M. Du Plessis, G. Niu, and M. Sugiyama, "Analysis of Learning from Positive and Unlabeled Data," in *NIPS*, 2014, pp. 703–711.
- [7] —, "Convex Formulation for Learning from Positive and Unlabeled Data," in *ICML*, 2015, pp. 1386–1394.
- [8] V. N. Vapnik, "An Overview of Statistical Learning Theory," *Neural Networks, IEEE Transactions on*, pp. 988–999, 1999.
- [9] C.-J. Hsieh, N. Natarajan, and I. S. Dhillon, "PU Learning for Matrix Completion," in *ICML*, 2015, pp. 2445–2453.
- [10] J. T. Zhou, S. J. Pan, Q. Mao, and I. W. Tsang, "Multi-View Positive and Unlabeled Learning," in *ACML*, 2012, pp. 555–570.
- [11] T. Sakai, M. C. d. Plessis, G. Niu, and M. Sugiyama, "Beyond the Low-Density Separation Principle: A Novel Approach to Semi-Supervised Learning," *arXiv preprint arXiv:1605.06955*, 2016.
- [12] X. Li, S. Y. Philip, B. Liu, and S.-K. Ng, "Positive Unlabeled Learning for Data Stream Classification," in *SDM*, vol. 9. SIAM, 2009, pp. 257–268.
- [13] M. N. Nguyen, X.-L. Li, and S.-K. Ng, "Positive Unlabeled Learning for Time Series Classification," in *IJCAI*, vol. 11, 2011, pp. 1421–1426.
- [14] J. Silva and R. Willett, "Hypergraph-Based Anomaly Detection of High-Dimensional Co-Occurrences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 563–569, 2009.
- [15] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially Supervised Classification of Text Documents," in *ICML*, vol. 2, 2002, pp. 387–394.
- [16] H. Yu, J. Han, and K. C.-C. Chang, "PEBL: Positive Example Based Learning for Web Page Classification Using SVM," in *International Conference on Knowledge Discovery and Data Mining*. ACM, 2002, pp. 239–248.
- [17] X. Li and B. Liu, "Learning to Classify Texts Using Positive and Unlabeled Data," in *IJCAI*, vol. 3, 2003, pp. 587–592.
- [18] H. Yu, "Single-Class Classification with Mapping Convergence," *Machine Learning*, vol. 61, no. 1-3, pp. 49–69, 2005.
- [19] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building Text Classifiers Using Positive and Unlabeled Examples," in *Third IEEE International Conference on Data Mining*. IEEE, 2003, pp. 179–186.
- [20] A. Skabar, "Single-class classifier learning using neural networks: An application to the prediction of mineral deposits," in *Machine Learning and Cybernetics, 2003 International Conference on*, vol. 4. IEEE, 2003, pp. 2127–2132.
- [21] G. Blanchard, G. Lee, and C. Scott, "Semi-Supervised Novelty Detection," *Journal of Machine Learning Research*, vol. 11, no. Nov, pp. 2973–3009, 2010.
- [22] J. Kittler, W. Christmas, T. De Campos, D. Windridge, F. Yan, J. Illingworth, and M. Osman, "Domain Anomaly Detection in Machine Perception: A System Architecture and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 845–859, 2014.
- [23] M. Markou and S. Singh, "Novelty Detection: A ReviewPart 1: Statistical Approaches," *Signal processing*, vol. 83, no. 12, pp. 2481–2497, 2003.
- [24] —, "Novelty Detection: A ReviewPart 2: Neural Network Based Approaches," *Signal processing*, vol. 83, no. 12, pp. 2499–2521, 2003.
- [25] M. Koppel and J. Schler, "Authorship Verification as a One-Class Classification Problem," in *ICML*, 2004, p. 62.
- [26] R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, and Q. Yang, "One-Class Collaborative Filtering," in *Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 502–511.
- [27] B. Schölkopf, R. Williamson, A. Smola, and J. Shawe-Taylor, "SV Estimation of a Distributions Support," *NIPS*, vol. 12, 1999.
- [28] D. M. Tax and R. P. Duin, "Support Vector Domain Description," *Pattern Recognition Letters*, vol. 20, no. 11, pp. 1191–1199, 1999.
- [29] M. M. Moya and D. R. Hush, "Network Constraints and Multi-Objective Optimization for One-Class Classification," *Neural Networks*, vol. 9, no. 3, pp. 463–474, 1996.
- [30] B. Schölkopf, J. C. Platt, and A. J. Smola, "Kernel Method for Percentile Feature Extraction," 2000.
- [31] D. M. Tax and R. P. Duin, "Uniform Object Generation for Optimizing One-Class Classifiers," *Journal of Machine Learning Research*, vol. 2, no. Dec, pp. 155–173, 2001.
- [32] C. Bennett and K. Campbell, "A Linear Programming Approach to Novelty Detection," *NIPS*, vol. 13, p. 395, 2001.
- [33] E. Pekalska, D. Tax, and R. Duin, "One-Class LP Classifiers for Dissimilarity Representations," in *NIPS*, 2002, pp. 761–768.
- [34] L. M. Manevitz and M. Yousef, "One-Class SVMs for Document Classification," *Journal of Machine Learning Research*, vol. 2, no. Dec, pp. 139–154, 2001.
- [35] D. M. Tax, *One-Class Classification*. TU Delft, Delft University of Technology, 2001.
- [36] D. M. Tax and R. P. Duin, "Combining One-Class Classifiers," in *International Workshop on Multiple Classifier Systems*. Springer, 2001, pp. 299–308.
- [37] A. D. Shieh and D. F. Kamm, "Ensembles of One Class Support Vector Machines," in *International Workshop on Multiple Classifier Systems*. Springer, 2009, pp. 181–190.
- [38] G. Ratsch, S. Mika, B. Schölkopf, and K.-R. Müller, "Constructing Boosting Algorithms from SVMs: An Application to One-Class Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1184–1199, 2002.
- [39] V. Jumutc and J. A. Suykens, "Multi-Class Supervised Novelty Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2510–2523, 2014.
- [40] G. Niu, M. C. d. Plessis, T. Sakai, and M. Sugiyama, "Theoretical Comparisons of Learning from Positive-Negative, Positive-Unlabeled, and Negative-Unlabeled Data," *arXiv preprint arXiv:1603.03130*, 2016.
- [41] S. S. Khan and M. G. Madden, "One-Class Classification: Taxonomy of Study and Review of Techniques," *The Knowledge Engineering Review*, vol. 29, no. 03, pp. 345–374, 2014.
- [42] B. M. Shahshahani and D. A. Landgrebe, "The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 5, pp. 1087–1095, 1994.
- [43] D. J. Miller and H. S. Uyar, "A Mixture of Experts Classifier with Learning Based on Both Labelled and Unlabelled Data," in *NIPS*, 1997, pp. 571–577.
- [44] T. Zhang and F. Oles, "The Value of Unlabeled Data for Classification Problems," in *ICML*, 2000, pp. 1191–1198.
- [45] O. Chapelle, B. Schölkopf, and A. Zien, "Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006)[Book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.
- [46] E. Sansone, A. Passerini, and F. G. B. D. Natale, "Clustering: Joint Classification and Clustering with Mixture of Factor Analysers," in *European Conference on Artificial Intelligence (ECAI)*, 2016, pp. 1089–1095.
- [47] O. Chapelle and A. Zien, "Semi-Supervised Classification by Low Density Separation," in *AISTATS*, 2005, pp. 57–64.
- [48] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.
- [49] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-Supervised Learning with Deep Generative Models," in *NIPS*, 2014, pp. 3581–3589.

- [50] Z.-H. Zhou and M. Li, "Semi-Supervised Learning by Disagreement," *Knowledge and Information Systems*, vol. 24, no. 3, pp. 415–439, 2010.
- [51] Y.-F. Li, I. W. Tsang, J. T. Kwok, and Z.-H. Zhou, "Convex and Scalable Weakly Labeled SVMs," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2151–2188, 2013.
- [52] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [53] P.-H. Chen, R.-E. Fan, and C.-J. Lin, "A Study on SMO-Type Decomposition Methods for Support Vector Machines," *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 893–908, 2006.
- [54] R. Fergus, Y. Weiss, and A. Torralba, "Semi-Supervised Learning in Gigantic Image Collections," in *NIPS*, 2009, pp. 522–530.
- [55] A. Talwalkar, S. Kumar, and H. Rowley, "Large-Scale Manifold Learning," in *CVPR*. IEEE, 2008, pp. 1–8.
- [56] Q. Da, Y. Yu, and Z.-H. Zhou, "Learning with Augmented Class by Exploiting Unlabeled Data," in *AAAI*, 2014, pp. 1760–1766.
- [57] M. Kearns, "Efficient Noise-Tolerant Learning from Statistical Queries," *Journal of the ACM (JACM)*, vol. 45, no. 6, pp. 983–1006, 1998.
- [58] F. De Comité, F. Denis, R. Gilleron, and F. Letouzey, "Positive and Unlabeled Examples Help Learning," in *International Conference on Algorithmic Learning Theory*. Springer, 1999, pp. 219–230.
- [59] F. Letouzey, F. Denis, and R. Gilleron, "Learning from Positive and Unlabeled Examples," in *International Conference on Algorithmic Learning Theory*. Springer, 2000, pp. 71–85.
- [60] J. He, Y. Zhang, X. Li, and Y. Wang, "Naive Bayes Classifier for Positive Unlabeled Learning with Uncertainty," in *SDM*. SIAM, 2010, pp. 361–372.
- [61] A. Jordan, "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes," pp. 841–848, 2002.
- [62] N. Aronszajn, "Theory of Reproducing Kernels," *Transactions of the American Mathematical Society*, pp. 337–404, 1950.
- [63] B. Schölkopf, R. Herbrich, and A. J. Smola, "A Generalized Representer Theorem," in *COLT*, 2001, pp. 416–426.
- [64] A. Argyriou and F. Dinuzzo, "A Unifying View of Representer Theorems," in *ICML*, 2014, pp. 748–756.
- [65] S. Boyd and L. Vandenberghe, *Convex Optimization*, 2004.
- [66] M. A. Hanson, "Invexity and the Kuhn–Tucker Theorem," *Journal of Mathematical Analysis and Applications*, vol. 236, no. 2, pp. 594–604, 1999.
- [67] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-dimensional Distribution," *Neural Computation*, pp. 1443–1471, 2001.
- [68] S. Mehrotra, "On the Implementation of a Primal-Dual Interior Point Method," *SIAM Journal on Optimization*, vol. 2, no. 4, pp. 575–601, 1992.
- [69] A. L. Yuille and A. Rangarajan, "The Concave-Convex Procedure (CCCP)," in *NIPS*, 2002, pp. 1033–1040.
- [70] W. S. Lee and B. Liu, "Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression," in *ICML*, vol. 3, 2003, pp. 448–455.
- [71] M. C. du Plessis, G. Niu, and M. Sugiyama, "Class-Prior Estimation for Learning from Positive and Unlabeled Data," in *ACML*, vol. 45, 2015.

TABLE 3

Comparative results (F-measure) on different small-scale datasets and on different values of hyperparameters using the linear kernel. 20% of positive examples are labeled, while the remaining are unlabeled.

Data	$\lambda = 0.0001$			$\lambda = 0.001$			$\lambda = 0.01$			$\lambda = 0.1$		
	[6]	[7]	USMO	[6]	[7]	USMO	[6]	[7]	USMO	[6]	[7]	USMO
<i>Australian</i>	69.8	67.9	68.3	69.9	69.6	67.7	71.5	69.6	69.3	68.9	70.1	70.0
<i>Clean1</i>	77.6	70.5	69.9	75.6	65.2	65.1	83.1	71.5	73.2	80.4	77.7	75.4
<i>Diabetes</i>	74.3	67.3	70.1	72.7	71.0	71.1	80.1	78.7	80.0	82.3	82.3	82.3
<i>Heart</i>	60.3	60.1	59.9	64.8	58.9	60.0	73.5	65.8	66.0	75.6	75.6	75.6
<i>Heart-statlog</i>	60.3	54.5	54.5	63.3	55.4	55.2	65.3	54.0	54.1	70.6	61.7	60.2
<i>House</i>	52.1	60.5	59.9	63.7	59.0	59.0	57.4	57.4	56.8	59.5	64.0	64.0
<i>House-votes</i>	53.1	59.6	59.6	55.8	60.0	60.2	46.9	52.9	57.7	36.8	60.2	62.0
<i>Ionosphere</i>	65.9	65.1	65.1	70.2	71.1	72.0	19.2	71.4	73.6	0.0	74.9	75.2
<i>Isotet</i>	80.1	93.7	93.7	83.9	93.3	93.5	83.2	95.3	94.7	84.5	98.5	95.8
<i>Krusk</i>	78.4	82.0	81.1	77.9	80.6	79.6	77.3	83.0	82.8	72.8	81.3	80.5
<i>Liverdisorders</i>	35.8	54.3	55.4	20.2	62.8	63.4	0.0	68.8	68.8	0.0	68.8	68.8
<i>Spectf</i>	53.4	73.5	73.5	51.0	72.9	72.3	0.0	80.3	80.6	0.0	81.1	81.1
Avg.	63.4±13.4	67.4±11.5	67.6±11.4	64.1±16.6	68.3±10.7	68.3±10.6	54.8±31.4	70.7±12.5	71.5±12.0	52.6±34.1	74.7±10.7	74.3±10.1

TABLE 4

Comparative results (training time in seconds) on different small-scale datasets and on different values of hyperparameters using the linear kernel. 20% of positive examples are labeled, while the remaining are unlabeled.

Data	$\lambda = 0.0001$			$\lambda = 0.001$			$\lambda = 0.01$			$\lambda = 0.1$		
	[6]	[7]	USMO	[6]	[7]	USMO	[6]	[7]	USMO	[6]	[7]	USMO
<i>Australian</i>	13.0	13.5	6.2	14.2	8.5	1.0	11.8	10.6	1.2	6.8	9.1	2.4
<i>Clean1</i>	26.0	8.9	7.2	20.1	7.4	4.8	8.5	6.8	4.9	6.7	6.2	2.3
<i>Diabetes</i>	27.5	35.3	7.7	47.9	29.3	3.6	27.6	15.6	2.9	5.1	22.0	2.2
<i>Heart</i>	3.1	2.2	1.2	1.0	1.5	0.5	3.1	1.1	0.7	3.2	5.2	0.6
<i>Heart-statlog</i>	9.5	6.1	0.9	3.7	2.2	1.1	0.8	1.2	0.6	1.0	1.5	0.7
<i>House</i>	1.5	0.9	0.5	2.0	1.3	0.5	1.4	0.8	0.4	0.6	0.7	0.3
<i>House-votes</i>	13.1	3.7	0.4	4.2	3.4	0.4	6.8	3.6	0.7	3.0	4.1	1.6
<i>Ionosphere</i>	12.9	6.5	1.7	16.4	8.2	1.7	4.0	6.1	1.5	5.4	6.4	0.9
<i>Isotet</i>	22.8	12.4	1.9	19.8	12.8	1.8	25.2	7.8	1.3	24.7	10.1	3.2
<i>Krusk</i>	575.4	540.2	35.3	546.1	510.8	40.0	252.9	323.3	43.4	233.5	153.4	22.8
<i>Liverdisorders</i>	7.3	7.0	1.5	6.9	5.6	0.9	4.2	5.6	1.0	3.9	7.4	0.5
<i>Spectf</i>	12.5	6.4	0.6	6.6	4.1	0.8	3.6	4.4	0.9	2.1	6.3	1.2
Avg.	60.4±162.4	53.6±153.5	5.4±9.8	57.4±154.4	49.6±145.4	4.8±11.2	29.2±71.0	32.2±91.8	5.0±12.2	24.7±66.1	19.4±42.6	3.2±6.2

TABLE 5

Comparative results (F-measure) on different small-scale datasets and on different values of hyperparameters using the Gaussian kernel (scale parameter equal to 1). 20% of positive examples are labeled, while the remaining are unlabeled.

Data	$\lambda = 0.0001$			$\lambda = 0.001$			$\lambda = 0.01$			$\lambda = 0.1$		
	[6]	[7]	USMO	[6]	[7]	USMO	[6]	[7]	USMO	[6]	[7]	USMO
<i>Australian</i>	64.4	64.2	64.7	66.6	70.7	70.4	67.1	57.6	59.4	66.7	0.0	0.0
<i>Clean1</i>	80.1	77.6	77.6	80.5	82.3	81.7	77.7	76.5	76.6	76.4	76.4	76.4
<i>Diabetes</i>	74.2	68.6	70.0	73.6	70.9	70.2	81.5	81.3	80.1	82.3	82.3	82.3
<i>Heart</i>	65.1	58.9	58.9	66.5	59.6	59.8	75.6	75.1	70.9	75.6	75.6	75.6
<i>Heart-statlog</i>	68.5	53.5	53.5	65.7	57.0	56.7	65.5	56.4	56.4	75.6	75.6	75.6
<i>House</i>	63.9	56.5	56.5	69.0	61.1	60.3	66.4	57.4	54.2	74.0	74.0	74.0
<i>House-votes</i>	36.8	61.2	61.2	46.6	59.2	59.2	41.3	56.1	57.5	0.0	71.8	71.8
<i>Ionosphere</i>	57.3	65.3	65.3	48.7	74.3	74.1	50.3	63.9	65.7	0.0	74.2	74.2
<i>Isotet</i>	73.3	75.3	75.3	73.8	75.7	75.7	71.4	0.0	76.4	71.4	75.9	75.9
<i>Krusk</i>	74.1	83.7	82.5	76.1	84.4	84.0	73.2	73.2	73.2	73.2	73.2	73.2
<i>Liverdisorders</i>	43.0	57.0	56.2	21.7	62.4	62.4	0.0	68.8	68.8	0.0	68.8	68.8
<i>Spectf</i>	39.2	74.0	73.7	47.6	72.9	72.6	0.0	81.1	81.1	0.0	81.1	81.1
Avg.	61.7±14.6	66.3±9.6	66.3±9.5	61.4±17.0	69.2±9.2	69.0±9.2	55.9±28.4	62.3±21.8	68.4±9.6	49.6±36.8	69.1±22.1	69.1±22.1

TABLE 6

Comparative results (training time in seconds) on different small-scale datasets and on different values of hyperparameters using the Gaussian kernel (scale parameter equal to 1). 20% of positive examples are labeled, while the remaining are unlabeled.

Data	$\lambda = 0.0001$			$\lambda = 0.001$			$\lambda = 0.01$			$\lambda = 0.1$		
	[6]	[7]	USMO	[6]	[7]	USMO	[6]	[7]	USMO	[6]	[7]	USMO
<i>Australian</i>	19.9	11.6	9.3	5.6	7.2	7.8	5.2	7.1	5.2	7.3	6.7	4.0
<i>Clean1</i>	1.5	3.6	23.1	1.7	2.6	17.0	2.3	3.0	8.7	1.3	2.4	7.1
<i>Diabetes</i>	16.7	12.8	6.7	25.8	10.9	4.5	9.1	7.9	4.8	4.0	11.8	11.4
<i>Heart</i>	3.2	1.1	0.6	0.5	0.9	0.7	0.5	0.6	0.5	0.5	1.1	0.7
<i>Heart-statlog</i>	1.5	1.1	1.7	0.6	1.1	0.7	0.4	0.6	0.6	0.5	0.9	0.8
<i>House</i>	0.4	0.7	1.5	0.4	0.8	1.3	0.3	0.5	0.6	0.4	0.4	0.5
<i>House-votes</i>	5.3	3.0	1.3	1.9	2.9	1.5	1.7	3.0	1.3	1.7	3.0	1.7
<i>Ionosphere</i>	3.4	1.8	1.7	1.2	2.1	1.6	2.1	1.7	2.6	1.3	1.9	1.9
<i>Isotet</i>	1.7	3.0	36.3	1.2	2.0	22.3	0.8	1.6	21.1	0.8	1.7	10.1
<i>Krusk</i>	148.6	244.7	136.7	118.1	154.7	123.8	410.9	107.6	133.1	162.9	154.0	178.3
<i>Liverdisorders</i>	2.1	1.7	1.8	1.9	1.6	1.0	1.0	2.8	0.9	1.1	2.3	1.3
<i>Spectf</i>	1.3	1.3	1.5	1.4	1.4	1.6	1.0	1.1	1.5	1.2	1.3	2.2
Avg.	17.1±41.9	23.9±69.7	18.5±38.8	13.4±33.8	15.7±43.9	15.3±34.9	36.3±118.0	11.5±30.4	15.1±37.6	15.3±46.6	15.6±43.7	18.3±50.5

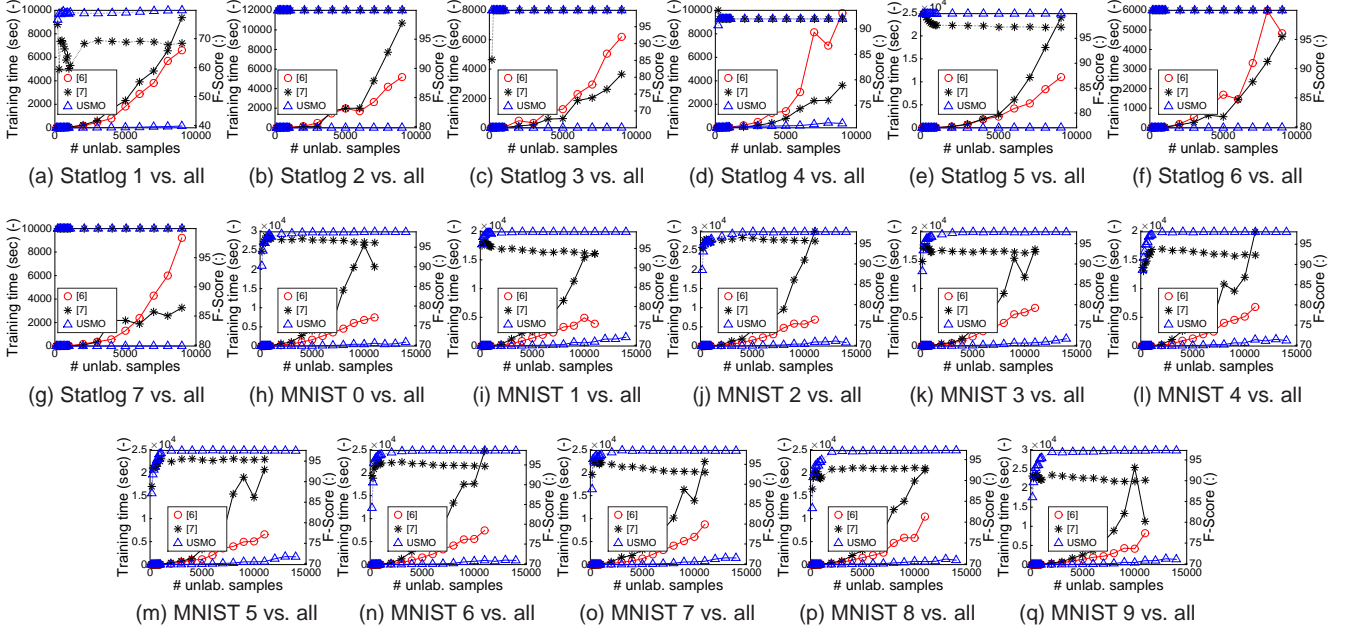


Fig. 4. Comparative results on (a)-(g) Statlog (shuttle) and (h)-(q) MNIST datasets using the linear kernel ($\lambda = 0.01$). Each plot shows the training time against different number of unlabeled samples as well as the generalization performance.

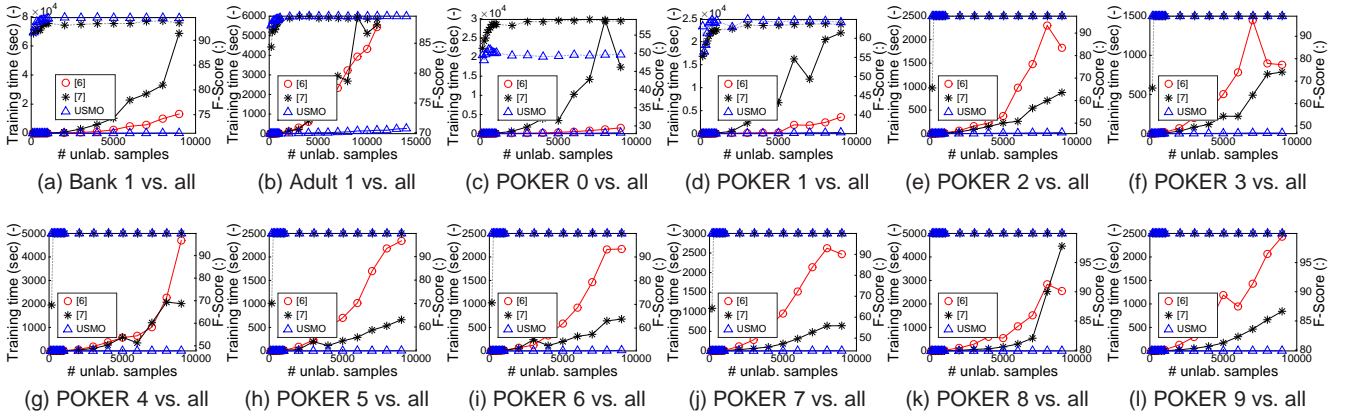


Fig. 5. Comparative results on (a) Bank-marketing, (b) Adult and (c)-(l) Poker-hand datasets using the linear kernel ($\lambda = 0.01$). Each plot shows the training time against different number of unlabeled samples as well as the generalization performance.